Yoga and Pilates Studios in Victoria: Finding a Market Need

Amanjit Gill

1. Introduction

1.1 Background

Yoga and pilates have soared in popularity in recent years. According to Roy Morgan Research, approximately 2.2 million Australians practise yoga, and 1.2 million practise pilates. Both of these activities are dominated by women; 77% of yoga devotees, and 90% of pilates participants, are women.

The growing market for yoga and pilates classes has driven growth in business opportunities for people working in health and fitness. However, with increased availability of teacher training courses, there is a risk that the market will become saturated with businesses offering classes in these disciplines.

Market saturation would see unsustainable competition among businesses, and a resultant downward pressure on class prices (and, therefore, revenue). Because of this, the long-term profitability of a yoga or pilates provider hinges upon locating its premises where there is an identified market need. Locations that are already well-serviced, or where the population demographics indicate a lack of interest in yoga or pilates, should be avoided.

1.2 Business question

If a new yoga or pilates studio were to open in Victoria, Australia, where should it be located?

1.3 Interested parties / stakeholders

This analysis is aimed at people in the health and fitness industry who are seeking to open a yoga or pilates studio in Victoria. It should also be of interest to people who invest in or provide finance to such businesses, as the insights in this analysis would aid in determining investment risk.

2. Data

Yoga Australia confirms our earlier assertion that yoga practice is dominated by women. It is also claimed that the majority of people who participate in yoga either hold tertiary educational qualifications or are currently studying at university. Because of this, data on demographics are required in order to confirm the link between educational attainment and participation in yoga and pilates.

In addition, it is well known that yoga and pilates (and, indeed, any exercise outside of the home) are discretionary expenditures. People living on low incomes such as the unemployed, or those on fixed pensions, may not be in a position to act upon an interest in yoga or pilates; these activities may be limited to people on higher incomes. Therefore, data

are required to determine the link, if one exists, between income, employment status and participation in yoga and pilates.

2.1 Data sources

2.1.1 Shapefile of local government areas (LGAs) in Victoria

Source: VIC LGA administrative boundaries files

This has been obtained from the Australian Government's publicly available data collection, and allows the creation of maps of Victoria that are segmented by LGA. There are also shapefiles available that allow segmentation by suburb, but segmentation by LGA is preferred. There are many hundreds of small suburbs in Victoria, and these are grouped into 79 incorporated LGAs.

Segmentation by LGA is preferred because some suburbs, particularly in Melbourne, are very small and are almost entirely composed of residential properties. Such suburbs would have very few (perhaps zero) yoga and pilates studios, and people living there would likely travel to another suburb in order to access yoga or pilates. Therefore, if data pertaining to studios are segmented by suburb, then the results may incorrectly indicate that there is no market for yoga or pilates in wholly residential suburbs. Using LGAs rather than suburbs captures those yoga and pilates participants who travel to adjacent suburbs within their own LGA.

2.1.2 General community profile (GCP) by LGA

Source: ABS census datapacks

The Australian Bureau of Statistics (ABS) has produced a number of 'datapacks' containing statistics that were collected during the last Australian census. The GCP datapack is the one that is most relevant to the problem at hand, because it contains data on the income and education demographics discussed earlier.

The GCP comprises 59 tables and 15522 columns of data about each LGA. Most of these columns are not required for the current analysis, and have been excluded. A small number of columns has been selected, because they pertain to sex, educational attainment, income and employment status - these are the demographic characteristics identified earlier as possibly having a deterministic effect on participation in yoga and pilates.

Where possible, statistics specifically pertaining to women are preferred over statistics pertaining to both sexes collectively; this is because the market for yoga and pilates is mostly comprised of women, so the exclusion of data about men yields a dataset that is more closely aligned with the principal participants in yoga and pilates.

The exception to this is the median family weekly income, which incorporates the incomes of both men and women. This has been included to account for the women who are

stay-at-home parents living with a partner, as these women would not be included in statistics related to work and income.

The statistics chosen from the 15222 columns in the GCP are shown in the table below.

Column Heading	Meaning	Notes
Tot_P_F	Total number of women in LGA	This is required for the calculation of percentages
Median_tot_fam_inc_weekly	Median total family income per week	
F_650_799_Tot F_800_999_Tot F_1000_1249_Tot F_1250_1499_Tot F_1500_1749_Tot F_1750_1999_Tot F_2000_2999_Tot F_3000_more_Tot	Number of women earning $650 to $799 per week / $800 to $999 per week / $1000 to $1249 per week / $1250 to $1499 per week / $1500 to $1749 per week / $1750 to $1999 per week / $2000 to $2999 per week / $3000 or more per week	These are to be summed and expressed as a percentage of the number of women living in the LGA
Percnt_Employment_to_populn_F	Percentage of women in the LGA who are employed	
F_PGrad_Deg_Total F_GradDip_and_GradCert_Total F_BachDeg_Total F_AdvDip_and_Dip_Total	Number of women with a postgraduate degree / graduate certificate or diploma / bachelor's degree / advanced diploma or diploma	These are to be summed and expressed as a percentage of the number of women living in the LGA

These columns have been extracted from their respective tables and merged into a single table containing 16 columns. Each row in the table pertains to one LGA, each of which has a unique code.

2.1.3 Australian Statistical Geography Standard (ADGS) for Victoria

Source: ABS ASGS datacubes

The GCP data include five-digit codes representing the LGAs, but not the names by which the LGAs are commonly known. An analysis that uses codes - rather than names - would be deficient in that the stakeholders would not readily recognise the LGAs by their codes. Therefore, a data source linking LGA codes with their corresponding names is required.

To this end, the ASGS has been found to be useful. This document contains a lot of information that is irrelevant to the current analysis, but it does identify the name belonging

to each LGA code, so this information has been extracted and added to the GCP dataset as another column.

2.1.4 Yoga/pilates studios by LGA

Source: Foursquare Places API

The number of yoga and pilates studios in each LGA has been obtained through the Foursquare API. The 'search' endpoint requires the provision of a location expressed as longitude and latitude; these coordinates have been obtained using a geocoder, then used in a call to the API.

The data from Foursquare have been used as an additional column in the table of GCP data. This table has then been subjected to cleaning, exploratory analysis and modelling in order to locate the LGAs that may have an unmet market need for new yoga and pilates studios.

2.1.5 Regional Population Growth by LGA

Source: ABS regional population growth datacubes

Knowledge of population growth is not required for this analysis. However, the ABS data on population growth include data on the land area of each LGA. This information has been extracted and used to estimate the radius of each LGA. This is required in order to supply the Foursquare API with a search radius within which to locate yoga and pilates studios in an LGA.

One limitation of the Foursquare API is that it doesn't search for venues within the bounds of a given locality (such as an LGA); rather, it takes the central coordinates of the locality and searches a circular area around that point. This makes the incorrect assumption that each LGA is circular in shape. However, it is expected that using a radius tailored to each LGA adequately compensates for this deficiency.

2.2 Final cleaned dataset

After the data are cleaned and the Foursquare API is accessed, the final dataset contains the following columns. There is one row per LGA, and there are 79 incorporated LGAs.

Feature Name	Feature Description
lga_code	unique identifier for the LGA
lga_name	name of the LGA
total_women	total number of women in the LGA
median_fam_inc_weekly	median family weekly income in the LGA
pcent_women_employed	percent of women in the LGA who are employed
pcent_women_high_income	percent of women in the LGA on a high income
pcent_women_tertiary_educ	percent of women in the LGA who are tertiary educated
studios_per_10000_women	number of studios in the LGA per 10000 women

3. Methodology

Of all the columns in the dataset, four are considered to be features that may be determinants of the demand for yoga and pilates in an LGA:

- median_fam_inc_weekly
- pcent_women_employed
- pcent_women_high_income
- pcent_women_tertiary_educ

These features have been subjected to two processes; exploratory analysis and the application of a machine learning model. These processes are described in more detail below.

3.1 Exploratory analysis

For the problem at hand, two simple approaches have been taken in analysing the features; the first involves the computation of descriptive statistics for each feature, and the other involves the computation of the Pearson correlation coefficient between each feature and the number of yoga and pilates studios in an LGA.

3.1.1 Descriptive statistics

The following descriptive statistics have been calculated for each feature, in order to gauge the likelihood of each feature being a determining factor in market demand for yoga and pilates studios.

- minimum value
- maximum value
- mean

- standard deviation
- 25th percentile
- 50th percentile i.e. median
- 75th percentile

These statistics have been used to compute the variability in each feature. Two measures for variability have been used:

- range (minimum value to maximum value)
- interquartile range (25th percentile to 75th percentile)

Features with greater variability are more likely to have an impact on the market for yoga and pilates studios than features with low variability. Therefore, computing the variability is one way to assess the suitability of the chosen features to the analysis at hand.

3.1.2 Pearson correlation

The Pearson correlation coefficient has been computed between each feature and the number of yoga studios per 10,000 women. Therefore, there are four computations of the correlation coefficient, between the following pairs of features:

- studios_per_10000_women and median_fam_inc_weekly
- studios_per_10000_women and pcent_women_employed
- studios_per_10000_women and pcent_women_high_income
- studios_per_10000_women and pcent_women_tertiary_educ

The magnitude of the correlation coefficient indicates the extent to which the feature is a determining factor of the prevalence of yoga and pilates studios in an LGA, and thus assists in determining the suitability of the feature's use in modelling.

3.2 Machine learning

As discussed earlier, the problem at hand is to assess which LGAs have a market need for more yoga and pilates studios. In order to make this determination, a machine learning model has been applied to the four relevant features identified earlier, as well as to the number of studios per 10,000 women.

Specifically, the LGAs in Victoria have been clustered using an unsupervised algorithm, k-Means. This algorithm works by computing the 'distance' between each LGA; not in terms of geography, but in terms of its features. The LGAs that are considered to be 'close together' form a cluster.

In applying a clustering algorithm, the expectation is that clusters are found which match the following profiles:

1. LGAs that are already well-serviced by studios, so there is no market need for more

2. LGAs with similar characteristics to the first cluster, but which are under-serviced and may have a market need for more studios

3. LGAs with little or no market need for more studios

It is expected that the first two clusters would comprise LGAs that are populated by women with high incomes and high levels of educational attainment, as these are the characteristics identified earlier as positively influencing participation in yoga and pilates. It is also expected that the third cluster would comprise LGAs that are populated by women with low incomes and low levels of educational attainment, as these characteristics are not considered to increase participation in yoga and pilates.

Assuming the successful formation of clusters meeting the profiles listed above, a final recommendation can be made that new yoga and pilates studios should be considered for LGAs in the second cluster; that is, LGAs with favourable demographics but which are under-serviced by studios.

4. Results

4.1 Exploratory analysis

4.1.1 Descriptive statistics

After computing the descriptive statistics for each feature, the middle 50% of the values (which represent the interquartile range) and the range (from the minimum to the maximum) have been computed. These two metrics give an indication of the variability in each feature.

Feature	Middle 50% of Values	Range (Min to Max)
median_fam_inc_weekly	1321.0 to 1814.5	1001.0 to 2765.0
pcent_women_employed	46.6 to 53.9	36.8 to 63.9
pcent_women_high_income	25.5 to 32.2	19.7 to 50.5
pcent_women_tertiary_educ	20.6 to 31.9	14.5 to 51.8

It has been found that the range for each feature is large, indicating a great difference in demographics between LGAs. The middle 50%, which represents the interquartile range, is also moderate to fairly large for each feature.

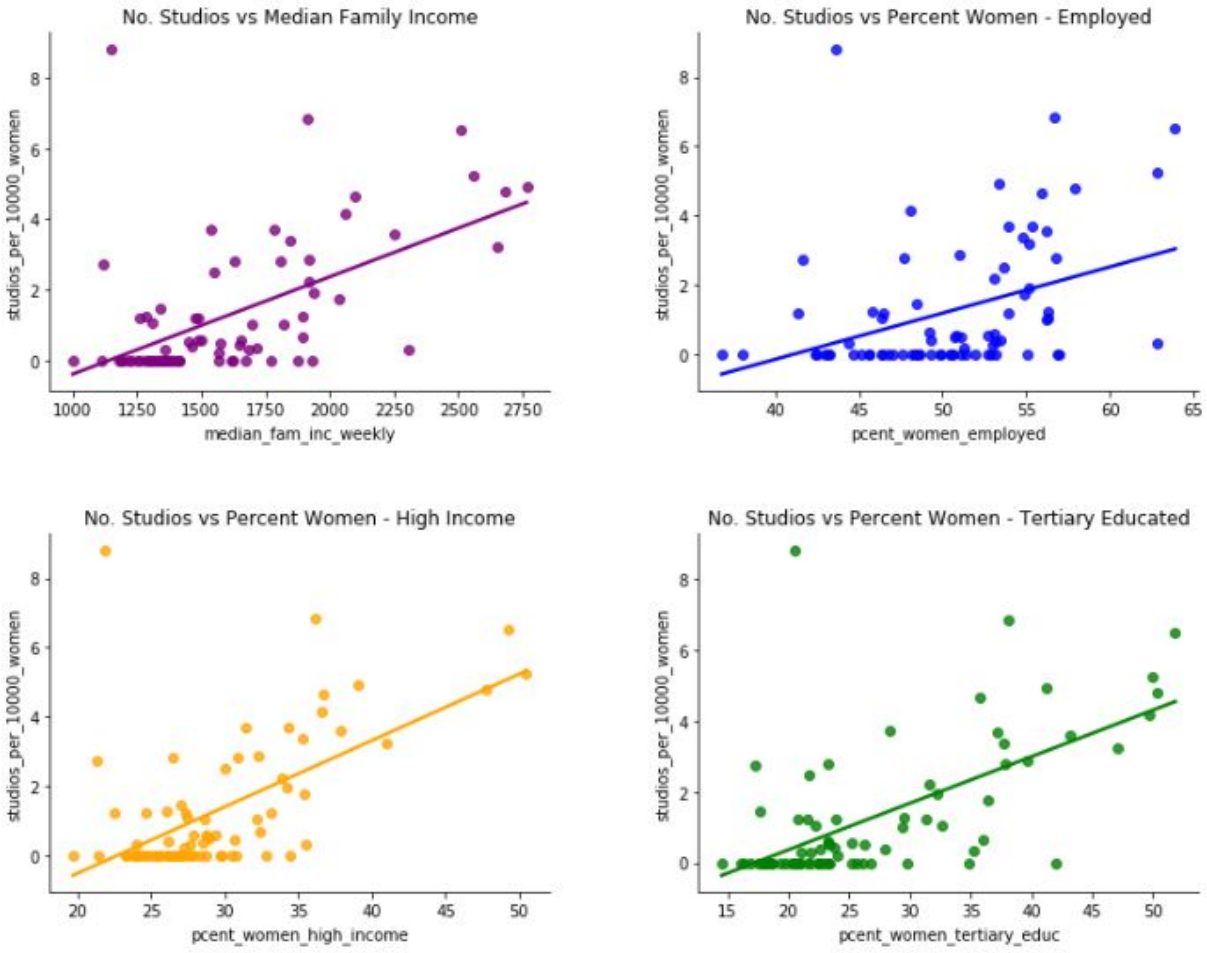
The high variability of all of the features suggests that these are appropriate features to have been selected for the problem at hand, as they may be determining factors in the number

and location of yoga and pilates studios. If the variability had been low, then these features would have been less likely to have an effect on the yoga and pilates market.

4.1.2 Pearson correlation

To confirm the extent to which the selected features are factors in the market for yoga and pilates, the correlation has been assessed between each feature and the number of yoga and pilates studios per 10,000 women.

Firstly, a scatter plot and regression line has been used to visualise the strength of the correlation for each feature. There are four scatter plots, one for each of the four features, as shown in the figure that follows.



It is apparent that there is a positive correlation between all four features and the number of studios per 10,000 women, but the magnitude of this correlation is not clear without a computation of the Pearson correlation coefficient. This computation has been completed and the results are as follows.

Feature	Pearson Correlation
median_fam_inc_weekly	0.566
pcent_women_employed	0.377
pcent_women_high_income	0.601
pcent_women_tertiary_educ	0.634

These results tell us that there is a moderate positive correlation between the following features and the prevalence of yoga and pilates studios in an LGA:

- median family weekly income
- percentage of women earning a high income
- percentage of women who are tertiary educated

In addition, it is apparent that there is only a weak correlation between the percentage of women who are employed and the prevalence of yoga and pilates studios in an LGA. Therefore, it is likely that while this feature may have some bearing on the market for yoga and pilates, it is not as strong a determinant as the other three features in the dataset.

This may be because many women in Australia who are in dual-income households do not do paid work; they are in an arrangement where a male partner is the breadwinner and the female is responsible for child-rearing. If the male partner's income is high enough, then the woman would be able to afford participation in yoga and pilates, thereby weakening the assumption that a woman must be earning her own money in order to participate in these activities.

Overall, the exploratory analysis has shown that three of the selected features are clearly appropriate for use in assessing where there may be a market need for yoga and pilates studios, while one of the features, the percentage of women who are employed in an LGA, has a smaller role to play but is still suitable for inclusion in modelling.

4.2 Machine learning

Having confirmed the suitability of the four selected features for use in modelling the problem at hand, a k-Means clustering model has been applied to the dataset, along with the fifth feature, the number of studios per 10,000 women. This unsupervised algorithm has grouped the LGAs into three clusters according to the LGAs' proximity to one another. As mentioned earlier, the distance between two LGAs is not geographical; rather, it is the extent to which the features of one LGA are similar to the features of another.

In order to determine the commonalities within each cluster, the mean of each feature within each cluster has been found. The results are shown below.

Cluster number	0	1	2
mean median_fam_inc_weekly	1333	1773	2387
mean pcent_women_employed	46.99	54.20	56.68
mean pcent_women_high_income	25.78	31.14	41.65
mean pcent_women_tertiary_educ	21.54	29.41	45.22
mean studios_per_10000_women	0.449	1.380	4.874

From these statistics, generalisations can be made about the LGAs in each cluster. These are summarised in the table below.

Cluster Number	Description
0	low family income; reasonably high employment of women; very low percentage of high income women; very low percentage of tertiary educated women; very few studios per 10,000 women
1	medium family income; high employment of women; low percentage of high income women; low percentage of tertiary educated women; few studios per 10,000 women
2	high family income; high employment of women; high percentage of high income women; high percentage of tertiary educated women; many studios per 10,000 women

From these observations, it is clear that the LGAs in Cluster 2 are already well-serviced by yoga and pilates studios. The LGAs in this cluster share wealthy demographics and high levels of educational attainment.

The LGAs in Cluster 0 has demographics that are not favourable for a potential new yoga or pilates studio. As discussed earlier, participation in yoga and pilates is related to wealth and educational attainment, and the LGAs in Cluster 0 do not meet these requirements. Therefore, a new studio would not be recommended to the LGAs in this cluster.

The LGAs in Cluster 1 share mixed demographics. While these LGAs feature families with a medium family income and high employment, the women in these LGAs earn low incomes and have lower levels of educational attainment than would be considered favourable for a potential new studio.

Overall, Cluster 2 is already well-serviced by yoga and pilates studios, Cluster 0 has wholly unfavourable demographics, and Cluster 1 has mixed demographics. Based on this information, no recommendation for a new yoga or pilates studio can be made with

confidence. Therefore, the clustering algorithm has been repeated; this time, though, there are four clusters instead of three, with the expectation that a cluster with favourable demographics that is under-serviced by studios will emerge.

The results of this four-cluster modelling are given below.

Cluster number	**0**	**1**	**2**	**3**
mean median_fam_inc_weekly	1535	2633	1935	1282
mean pcent_women_employed	52.27	58.64	54.36	45.02
mean pcent_women_high_income	28.23	45.50	34.19	24.81
mean pcent_women_tertiary_educ	23.57	48.08	36.13	20.85
mean studios_per_10000_women	0.430	4.928	2.634	0.622

As before, these statistics can be used to make generalisations about each cluster. These generalisations are described below.

Cluster Number	**Description**
0	low family income; high employment of women; very low percentage of high income women; very low percentage of tertiary educated women; very few studios per 10,000 women
1	high family income; high employment of women; high percentage of high income women; high percentage of tertiary educated women; many studios per 10,000 women
2	moderately high family income; high employment of women; low-med percentage of high income women; low-med percentage of tertiary educated women; some studios per 10,000 women
3	very low family income; moderate employment of women; very low percentage of high income women; very low percentage of tertiary educated women; very few studios per 10,000 women

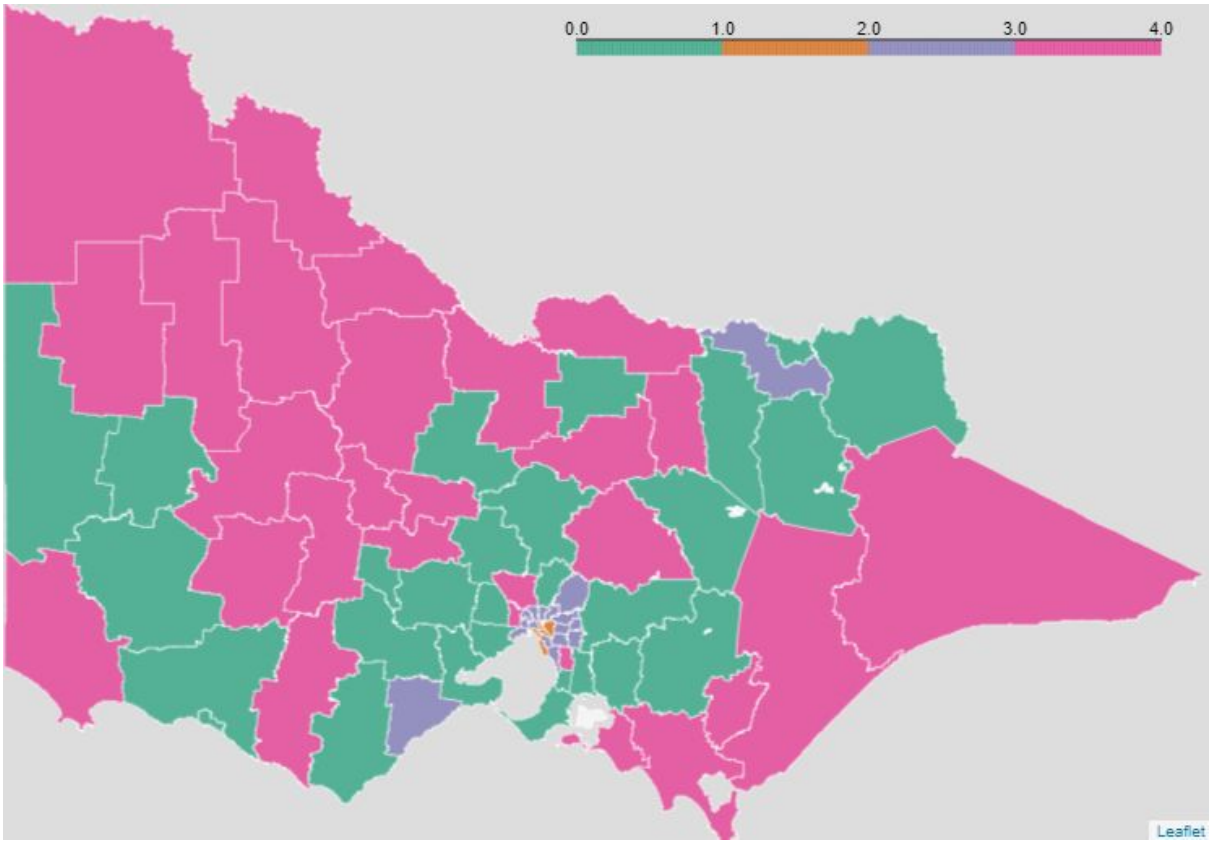
From these observations, it is clear that Cluster 1 comprises LGAs that are already well-serviced by yoga and pilates studios. By contrast, Clusters 0 and 3 share similar demographics - low incomes and low levels of educational attainment - that render them unsuitable for a new studio. Therefore, a new studio would not be recommended for Clusters 0, 1 and 3.

The LGAs in Cluster 2 share favourable demographics - these LGAs are not as wealthy as those in Cluster 1, but the women in these LGAs enjoy moderately high family incomes, even if their personal incomes and levels of educational attainment are not as high as the women who live within Cluster 1.

Cluster 2 is not as well-serviced by yoga and pilates studios as Cluster 1. Therefore, we can recommend with some confidence that there is a market need in Cluster 2 for a new studio.

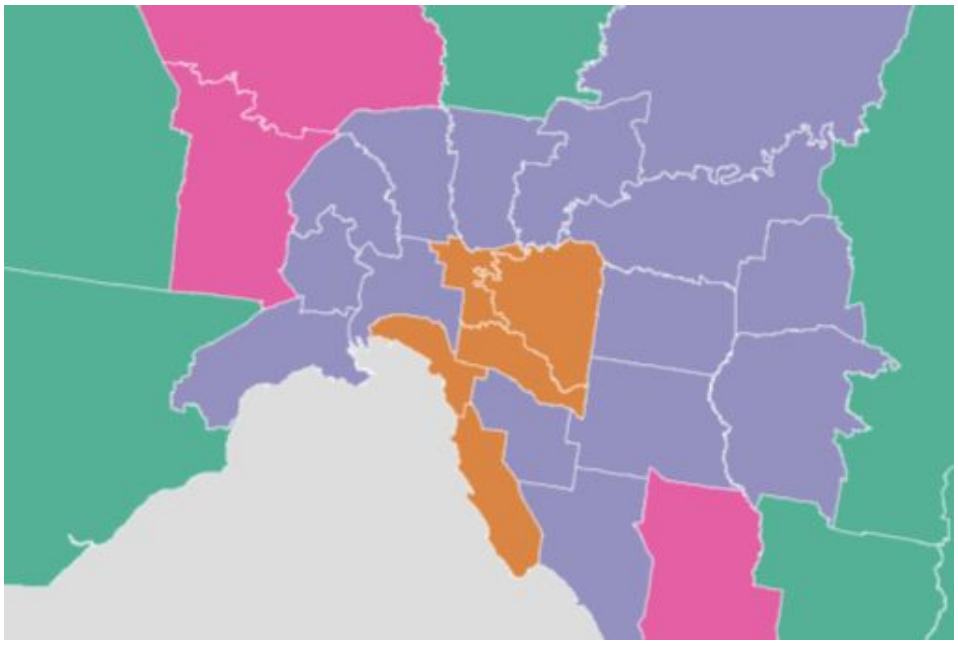
5. Discussion

It has been established that the greatest market need for new yoga and pilates studios exists in the LGAs belonging to Cluster 2. In order to gain insight into any geographic commonalities that these LGAs share, all four clusters are shown on the map below. Note that the white gaps in the map correspond to small areas of Victoria that are not incorporated LGAs and have thus been excluded from this analysis.



It is apparent that regional Victoria is dominated by Clusters 0 and 3 (green and magenta). This means that these regions, which are located far away from the capital city, Melbourne, mostly share demographics that are not favourable for people wishing to open businesses offering services in yoga and pilates.

Upon closer inspection, it is evident that all of Cluster 1 (brown) is located in central Melbourne, and it is surrounded by LGAs belonging to Cluster 2 (blue). A closer view of Melbourne is provided in the graphic that follows.



This indicates that the wealthiest Victorians - those who already have the best access to yoga and pilates studios - live in central Melbourne, and they are surrounded by communities that may have an unmet market need for more studios. These LGAs belong to Cluster 2. With just two exceptions, all of the LGAs in Cluster 2 are located in the vicinity of Melbourne, surrounding Cluster 1.

Therefore, it is recommended that a businessperson wishing to open a new yoga or pilates studio select a location from within Cluster 2, which comprises the following LGAs:

Banyule	Knox	Moonee Valley
Darebin	Manningham	Moreland
Glen Eira	Maribyrnong	Nillumbik
Hobsons Bay	Maroondah	Surf Coast
Indigo	Melbourne	Whitehorse
Kingston	Monash	

6. Conclusion

A relationship between wealth, education and participation in yoga and pilates has been confirmed. This has been used to group the local government areas (LGAs) of Victoria into

clusters, in order to locate areas where there is an unmet market need for yoga and pilates studios.

The machine learning algorithm that has been used to form these clusters has successfully identified a number of LGAs with favourable demographics that are under-serviced by yoga and pilates businesses. It is recommended that businesses wishing to provide services in yoga and pilates should consider locating their premises within these LGAs.