The Battle of Neighborhoods

Jason Kenner

1. Introduction

1.1 Background. Melbourne is a diverse, multicultural city that has been growing quickly. The demographics are changing with population growth. There is a strong food and entertainment culture including a great café and bar scene. There are many opportunities available for hardworking entrepreneurs who are savvy investors, particularly taking into account location when starting up a new hospitality venture. Melbourne has historically had working class and middle class suburbs. The working class suburbs had many pubs and bars, while the middle class had fewer.

1.2 Problem. Over the past 30 years, inner working class suburbs have become gentrified. Therefore, the assumptions about suburbs and behaviours are less easily predicted.
This project aims to see if data science can provide suggestions of suburbs that would be good candidates for new bars buy clustering postcodes based on demographic data and then comparing the number of bars in like clusters.

1.3 Interest. The client wants to open a bar in the inner suburbs of Melbourne and would like to know which suburbs have fewer bars, and also demographic data that suggests bars are well frequented with the people who reside there.

2. Data

2.1 Data sources. Data was scraped from several sources, and also obtained from Foursquare. Firstly, postcode data including latitude and longitude was obtained from https://www.matthewproctor.com/australian_postcodes as a csv file. It appeared as such in a data frame:

	id	postcode	locality	state	long	lat	dc	type	status	sa3	sa3name	sa4	sa4name	region
0	230	200	ANU	ACT	0.000000	0.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	R1
1	21820	200	Australian National University	ACT	149.118900	-35.277700	NaN	NaN	Added 19-Jan-2020	NaN	NaN	NaN	NaN	R1
2	232	800	DARWIN	NT	130.836680	-12.458684	NaN	NaN	Updated 6-Feb-2020	70101.0	Darwin City	701.0	Darwin	R1
3	233	801	DARWIN	NT	130.836680	-12.458684	NaN	NaN	Updated 25-Mar-2020 SA3	70101.0	Darwin City	701.0	Darwin	R1
4	234	804	PARAP	NT	130.873315	-12.428017	NaN	NaN	Updated 25-Mar-2020 SA3	70102.0	Darwin Suburbs	701.0	Darwin	R1

Also, demographic data was obtained from the Australian Bureau of Statistics http://stat.data.abs.gov.au/index.aspx# also as a csv file. It was a much more complex set of data including many rows with the same postcode:

	POA	Postal Area Code	SEIFAINDEXTYPE	Index Type	SEIFA_MEASURE	Measure	TIME	Time	Value	Flag Codes	Flags
0	800	800	IEO	Index of Education and Occupation	SCORE	Score	2016.0	2016.0	1089.0	NaN	NaN
1	800	800	IEO	Index of Education and Occupation	RWAR	Rank within Australia	2016.0	2016.0	2287.0	NaN	NaN
2	800	800	IEO	Index of Education and Occupation	RWAD	Rank within Australia - Decile	2016.0	2016.0	9.0	NaN	NaN
3	800	800	IEO	Index of Education and Occupation	RWAP	Rank within Australia - Percentile	2016.0	2016.0	87.0	NaN	NaN
4	800	800	IEO	Index of Education and Occupation	RWSR	Rank within State or Territory	2016.0	2016.0	33.0	NaN	NaN

Finally, location data was obtained from Foursquare https://developer.foursquare.com as a json file.

	Locality	Locality lat	Locality long	Bar	Bar lat	Bar long	Bar category
0	MELBOURNE	-37.817403	144.956776	Dikstein's Corner Bar	-37.816189	144.960353	Bar
1	MELBOURNE	-37.817403	144.956776	The Irish Times	-37.816135	144.960563	Bar
2	MELBOURNE	-37.817403	144.956776	MoVida Terraza	-37.814688	144.958567	Bar
3	MELBOURNE	-37.817403	144.956776	Bonnie Coffee Brewers	-37.818153	144.957636	Coffee Shop
4	MELBOURNE	-37.817403	144.956776	Patricia Coffee Brewers	-37.814598	144.958350	Coffee Shop
5	MELBOURNE	-37.817403	144.956776	The Lui Bar	-37.819067	144.957739	Cocktail Bar
6	MELBOURNE	-37.817403	144.956776	Shamble Coffee Brewers	-37.816056	144.960779	Café
7	MELBOURNE	-37.817403	144.956776	The Deck	-37.820254	144.957515	Bar
8	MELBOURNE	-37.817403	144.956776	Syracuse	-37.816207	144.960253	Restaurant
9	MELBOURNE	-37.817403	144.956776	Saint & Rogue	-37.817512	144.955491	Bar

2.2 Data cleaning. The postcode and demographic data required cleaning, including removal of unwanted columns and rows, renaming headings and merging. The demographic data was firstly made into three new data frames, one for each demographic data point, then those data frames combined, and finally merged with the postcode data. The result included postcode, latitude, longitude and three demographic measures: Index of Education and Occupation, Index of Economic Resources, and Index of Social Advantage and Disadvantage for each suburb. The dataframe was then reduced to only include suburbs within a 5km radius of the Melbourne City Centre using a csv file obtained from FreeMapTools: https://www.freemaptools.com/find-australian-postcodes-inside-radius.htm. The final data frame of inner suburb data was created by calling the postcode subset data frame using .isin:

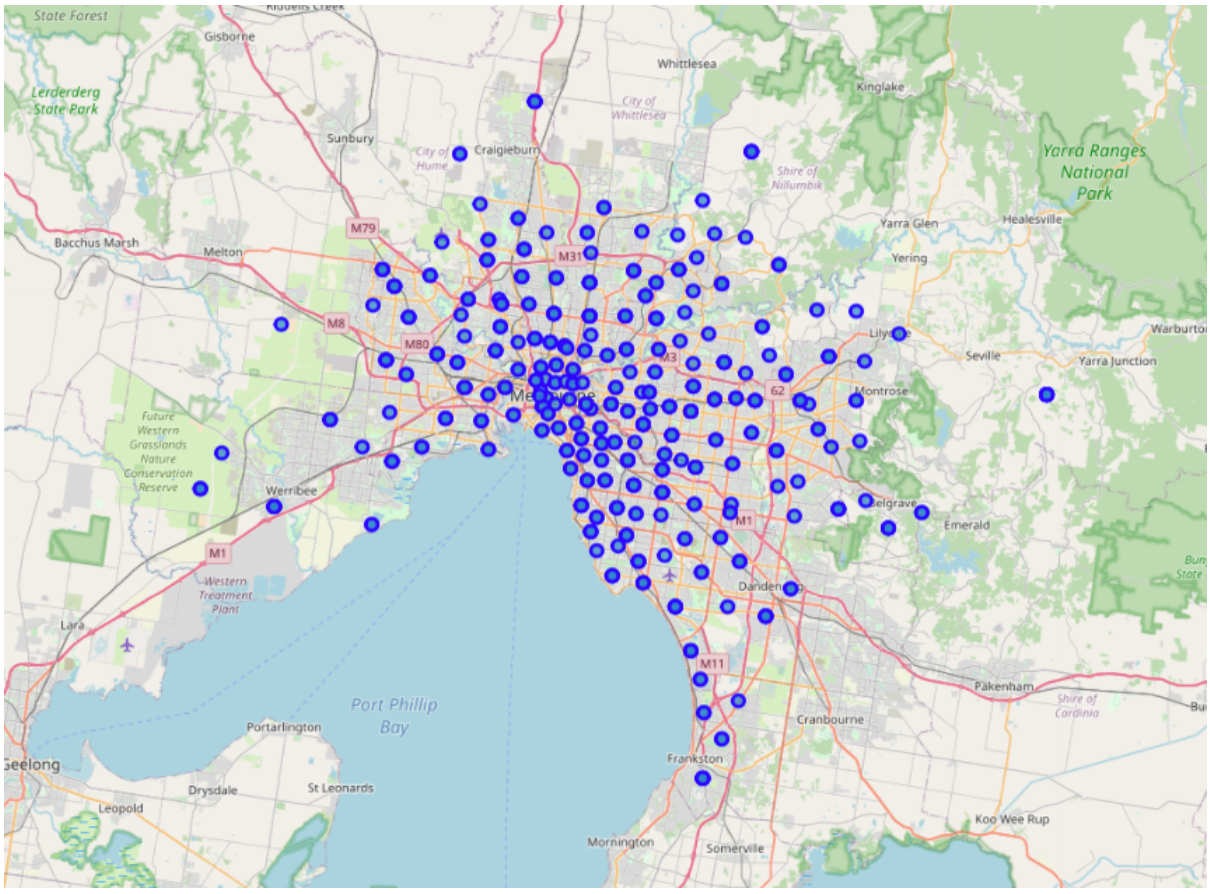
	postcode	locality	state	long	lat	ieo rank	ier rank	irsad rank
0	3000	MELBOURNE	VIC	144.956776	-37.817403	10.0	1.0	8.0
1	3002	EAST MELBOURNE	VIC	144.982207	-37.818517	10.0	4.0	10.0
2	3003	WEST MELBOURNE	VIC	144.949592	-37.810871	10.0	1.0	10.0
3	3004	MELBOURNE	VIC	144.970161	-37.844246	10.0	2.0	10.0
4	3004	ST KILDA ROAD CENTRAL	VIC	144.970161	-37.844246	10.0	2.0	10.0

The Foursquare data was also grouped by suburb into a new data frame and combined with the demographic data for each suburb for the clustering analysis allowing a single data frame that also included total number of bars for each suburb:

	postcode	locality	long	lat	ieo rank	ier rank	irsad rank	Bars
0	3000	MELBOURNE	144.956776	-37.817403	10.0	1.0	8.0	54.0
1	3002	EAST MELBOURNE	144.982207	-37.818517	10.0	4.0	10.0	6.0
2	3003	WEST MELBOURNE	144.949592	-37.810871	10.0	1.0	10.0	6.0
3	3004	MELBOURNE	144.970161	-37.844246	10.0	2.0	10.0	54.0
4	3005	WORLD TRADE CENTRE	144.950858	-37.824608	10.0	1.0	10.0	11.0

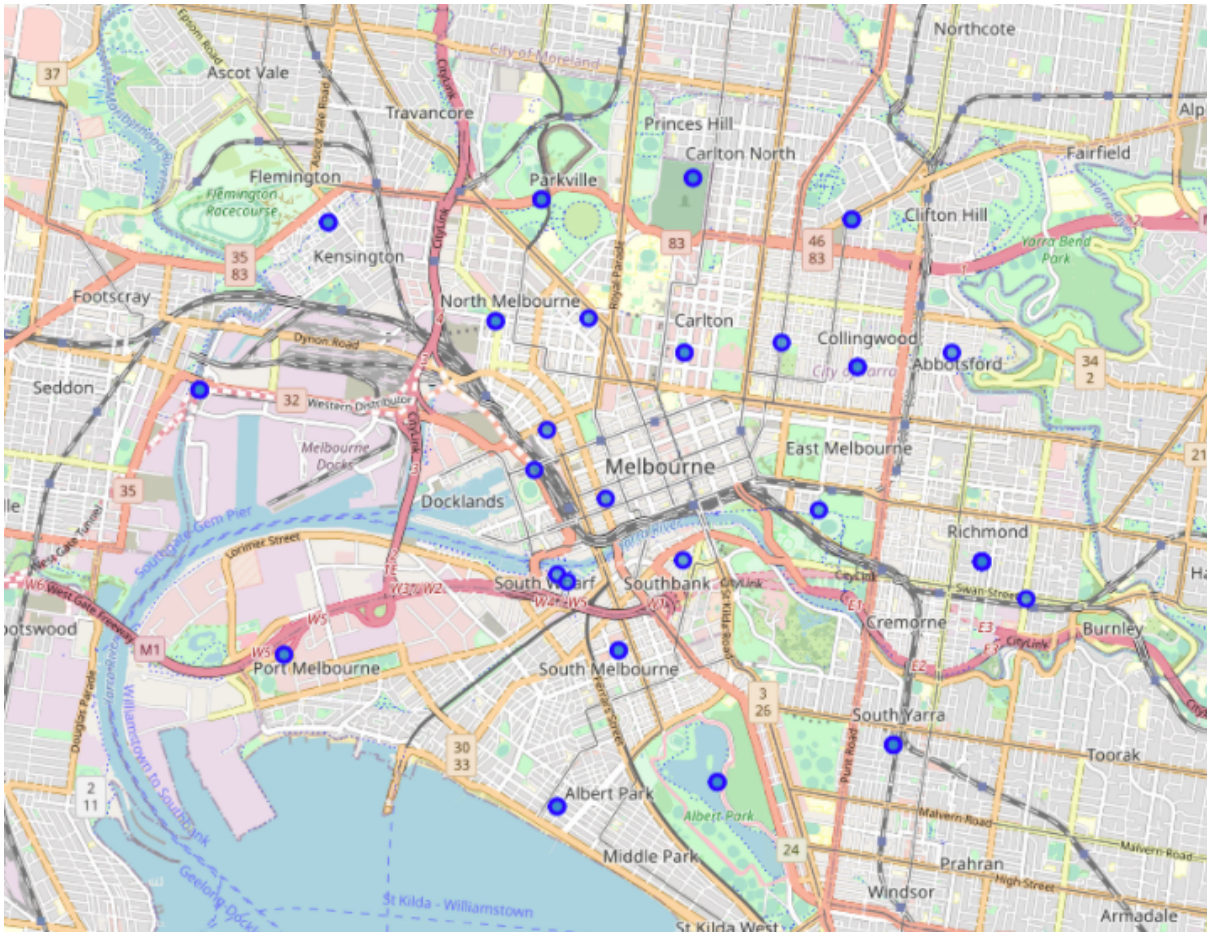
3. Methodology

3.1 Check and visualise the postcode data. The initial data collection and preparation required combining data from two different sources as described in section 2 above. Once that data frame was complete, I checked to see if the coordinates looked good on a map of Melbourne metropolitan area:



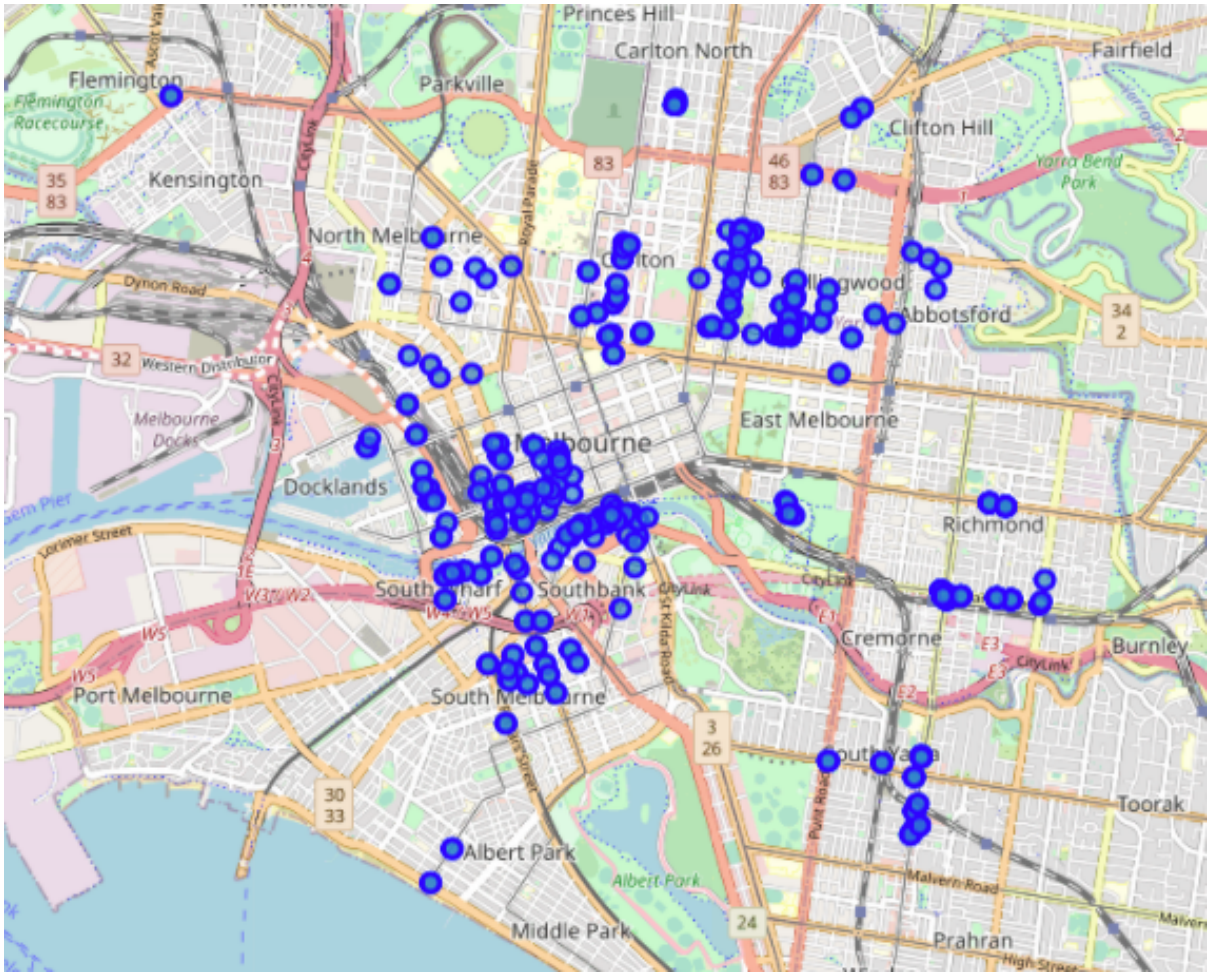
First map with all Melbourne metropolitan suburbs.

This looked fine; however, the project only requires data within 5kms of Melbourne. To reduce the data frame, I needed first to obtain a list of only the postcodes within the inner 5km radius. This was obtained as a csv file from FreeMapTools and then saved to a GitHub to be called from the notebook. I was then able to create a new data frame from only those postcodes. This was confirmed by plotting them against a map of inner Melbourne:



Second map with only inner Melbourne suburbs.

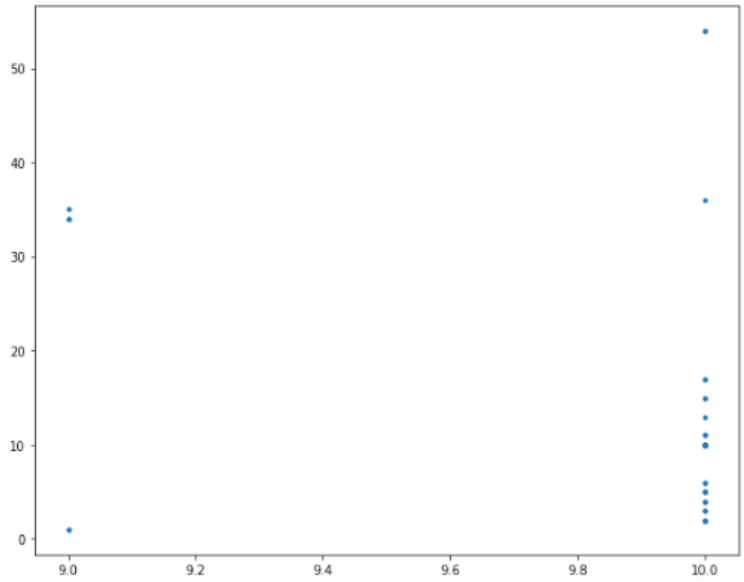
3.2 Obtain venue data from Fourquare. When obtaining Foursquare data, I include the categoryId field to ensure that I only received venues classified as a bar for all of the suburbs identified as residing within 5kms of Melbourne. These were also plotted on a map of Melbourne:



Map of the bars within 5kms of Melbourne.

The purpose of the project was to use demographic data to see if like suburbs contained varying numbers of bars. The idea being that if a suburb with similar demographic data to another has significantly fewer bars, then opening a new bar in that suburb would have a higher chance of success than opening one in the suburb that had more bars by comparison. So, the next step was to group the venue data by suburb so that the total number of bars per suburb could be used as a data point. Creation of the data frame is described in section 2 above.

3.3 Group and analyse data. I then plotted each of the three demographic rankings against number of bars to see if there were some obvious trends. Fist was Index of Education and Occupation (IEO):

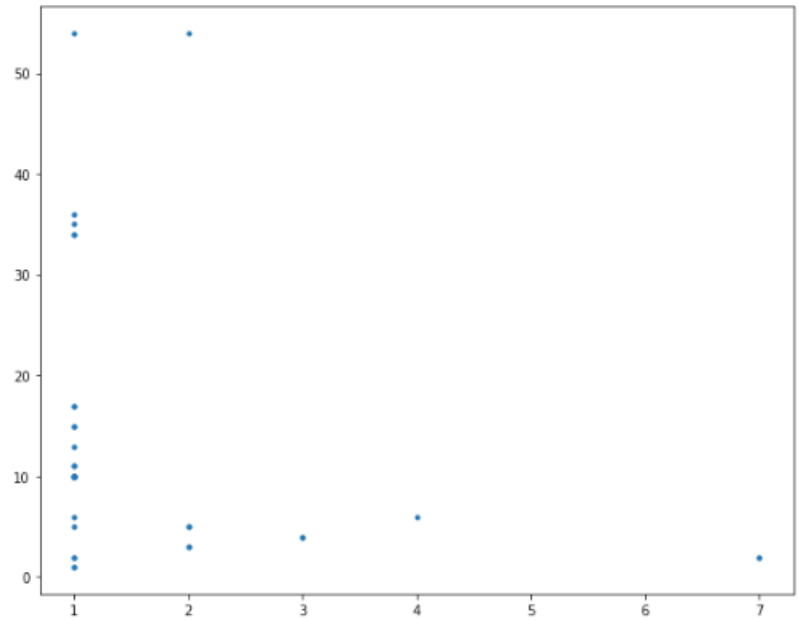


Plot of IEO against total bars per suburb.

According to the ABS, IEO: A low score indicates relatively lower education and occupation status of people in the area in general. A high score indicates relatively higher education and occupation status of people in the area in general.

All IEO values in the plot are 9 or 10 out of 10. This suggest all suburbs represented in the data frame have relatively high education and occupation status. Because they are all high, it's not going to have a significant contribution to the clustering analysis. Nonetheless, it is an interesting observation of people who live close to the CBD.

The next plot observed Index of Economic Resources IER:

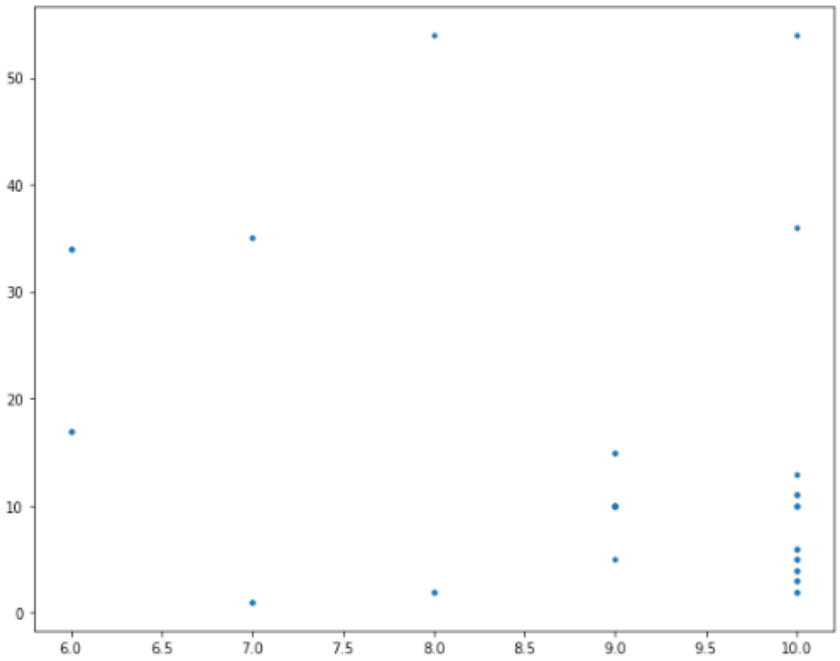


Plot of IER against total bars per suburb.

According to the ABS, IER: A low score indicates a relative lack of access to economic resources in general. While a high score indicates relatively greater access to economic resources in general.

There is more variation in the IER score indicating more variation in access to economic resources in the inner suburbs. The plot demonstrates that there are more bars in suburbs with low IER scores, indicates a relative lack of access to economic resources in general.

The third plot observed Index of Social Advantage and Disadvantage (IRSAD):

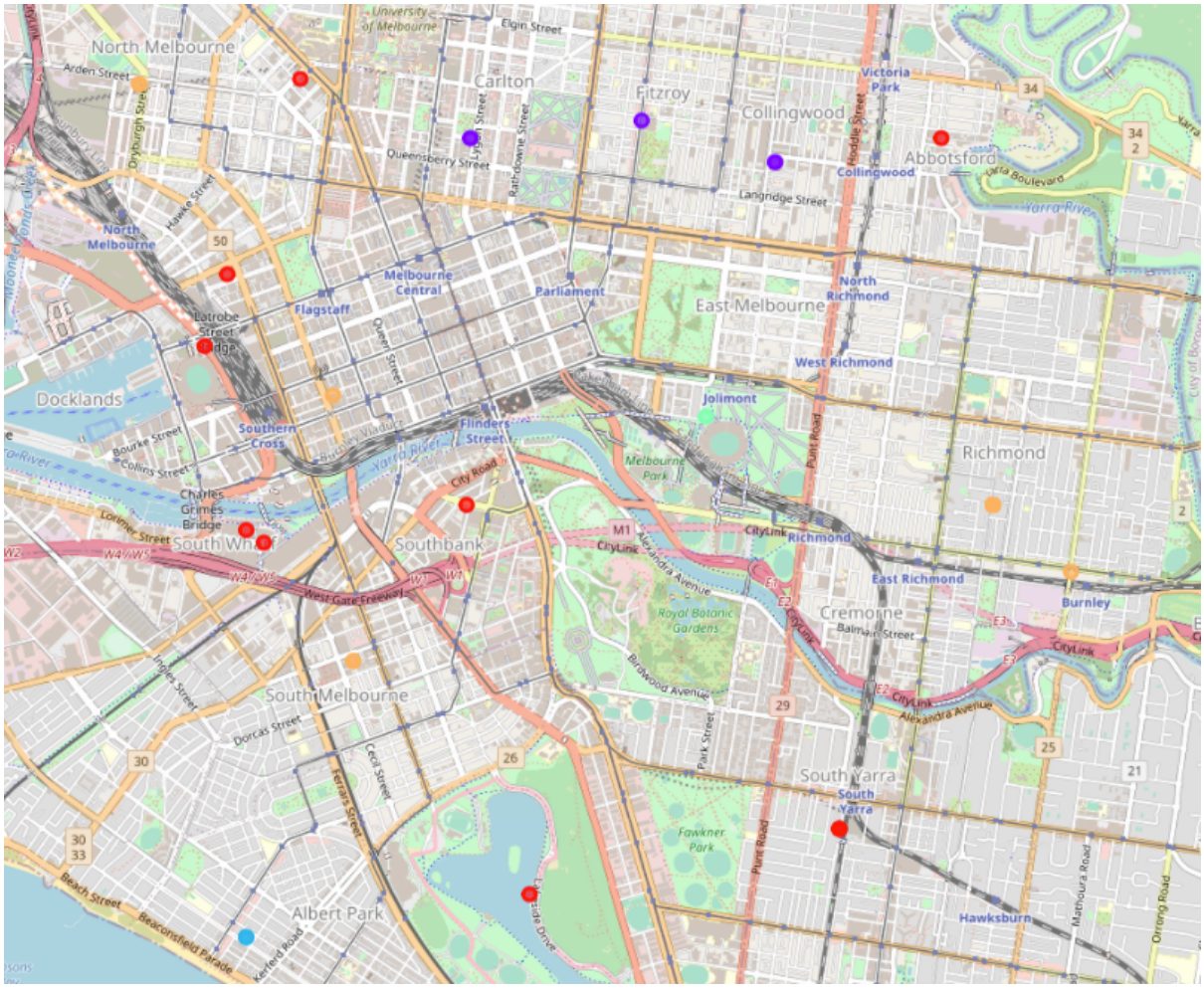


Plot of IRSAD against total bars per suburb.

According to the ABS, IRSAD: A low score indicates relatively greater disadvantage and a lack of advantage in general. A high score indicates a relative lack of disadvantage and greater advantage in general.

The plot shows a range of scores between 6-10 out of 10 with somewhat more bars in suburbs with higher scores indicating a relative lack of disadvantage and greater advantage in general.

3.4 Cluster and analyse clusters. I felt the plots suggested clustering on these three rankings could provide some helpful suggestions for choosing a location for a new bar and performed k-means clustering on the data. I used 5 clusters and plotted them on a map:



Map of demographic clusters.

The final step was to sort the grouped data frame by cluster and by number of bars to see if there are some demographic clusters that have significant differences in the number of bars in the individual suburbs. The sorted data frame is described in section 2 above.

4. Results

Of the five clusters, three clusters did have significant differences. These were:

4.1 Cluster 0. Cluster 0 contained 13 suburbs. These suburbs all had high IEO scores (10/10), high IRSAD scores (10/10) and low IER scores (1 or 2/10). The number of bars per suburb ranged from 3 to 54. While the two highest scoring suburbs are the two main city suburbs, the others still ranged from 3 to 13. The two suburbs with only 3 bars were Carlton North and Princes Hill.

4.2 Cluster 1. Cluster 1 contained 7 suburbs. These suburbs all had high IEO scores (9 or 10/10), middle IRSAD scores (6 or 7/10) and low IER scores (1/10). The number of bars per suburb ranged from 1 to 35. The two suburbs with scores of 1 were Flemington and Kensington.

4.3 Cluster 4. Cluster 4 contained 13 suburbs. These suburbs all had high IEO scores (10/10), relatively high IRSAD scores (8 or 9/10) and low IER scores (1/10). The number of bars per suburb ranged from 2 to 54. While the highest scoring suburb was in the CBD, the others still ranged from 2 to 15. The two suburbs with scores of 2 were Hotham Hill and North Melbourne.

5. Discussion

5.1 Using demographic data to select bar location. The data analysis revealed that within some clusters there is significant variation in the number of bars. While there may be other factors not included in the analysis, this project has provided some useful insights for an entrepreneur interested in opening a bar in inner Melbourne.

5.2 Recommendation. Based on the analysis, there are 6 suburbs that make good candidates for opening a new bar. These are: Carlton North, Princes Hill, Flemington, Kensington, Hotham Hill and North Melbourne.

6. Conclusion

This analysis made use of data visualisation and clustering to analyse demographic and venue data the is freely available to make business decisions. The insights provided through the analysis suggest that this form of analysis can inform decision related to location of a new business such as a bar. It would seem that similar methods could be applied to a range of site specific business also.