Improving Family-friendly Neighbourhoods in Melbourne, AU

Israel Tiomno
25th May 2020

Introduction

The local government of Melbourne in Australia is interested in obtaining information about the places and facilities that are suitable for families with children under 15 years old. The goal is to develop local communities' public areas for families with children.

This study aims to find which suburbs need resources and specific projects to improve those areas for the segment of the population being targeted.

Data Description

This study used demographic data from the Australian Bureau of Statistics (ABS) to compile information about families per suburb and the API of Foursquare to gather the venues of each locality.

The first dataset downloaded from ABS has aggregations of family composition and other statistics grouped by the suburbs of the State of Victoria in Australia. From its several columns, only four were selected for being relevant to this study: `SSC_CODE_2016` gives the suburb code (State Suburb Code assigned by the ABS), `CF_ChU15_a_Total_F` the number of couple families with children under 15 years, `OPF_ChU15_a_Total_F` the number of one-parent families with children under 15 years, and `Total_F` the total of families. An excerpt of this dataset can be seen in (*Table 1*)

Table 1. Excerpt of ABS Family Composition dataset.

SSC_CODE_2016	...	CF_ChU15_a_Total_F	...	OPF_ChU15_a_Total_F	...	Total_F	...
...	...	...	...	...	...	...	...
SSC20002	...	353	...	58	...	1886	...
SSC20003	...	356	...	39	...	1034	...
...	...	...	...	...	...		

From the above dataset, the ratio of families with children from the total of families can be calculated with the formula:

```
(CF_ChU15_a_Total_F + OPF_ChU15_a_Total_F) / Total_F
```

The Victorian government provided the dataset `CG_SSC_2016_SA4_2016.csv` which put together the correspondence between SSC and SA4. The later code represents larger statistic areas than suburbs defined by the ABS. Those that have 'Melbourne' as part of its name in the field `SA4_NAME_2016` represent the areas of the city of Melbourne with their corresponding suburbs. This dataset was used to filter out any suburb not located in the Melbourne area. An excerpt of this dataset can be seen in (*Table 2*)

Table 2. Excerpt of ABS Correspondence between SSC and SA4 areas dataset

SSC_CODE_2016	SSC_NAME_2016	SA4_CODE_2016	SA4_NAME_2016	RATIO	PERCENTAGE
20172	Bayswater (Vic.)	211	Melbourne - Outer East	1	100
20173	Bayswater North	211	Melbourne - Outer East	1	100
20174	Beaconsfield (Vic.)	212	Melbourne - South East	1	100
20175	Beaconsfield Upper	212	Melbourne - South East	1	100
20176	Bealiba	201	Ballarat	1	100
20177	Bearii	216	Shepparton	1	100
20178	Bears Lagoon	202	Bendigo	1	100

This specific study was focused on the suburbs of Melbourne – Inner, the SA4 area most dense populated in the city. Each suburb geolocation was used to retrieve the venues suitable for families with children accessing the Foursquare API. To define the suitability of places, a list of 50 categories were selected form the categories list that Foursquare publishes in its website https://developer.foursquare.com/docs/resources/categories.

The list of venues per suburb were merged to the ratio of families to obtain a final dataset ready to be analysed. A clear approach was to create segments by grouping suburbs with similar number of venues, types of venues and number of families.

Methodology

The goal of this study is to identify a pattern in the suburbs of Melbourne based on type of venues and number of families with children. Since there was no previous data that classify the suburbs in such a way, a clustering machine learning algorithm is the most appropriate. This was concluded as the data at hand showed no way to label the suburbs with the necessary characteristics, thus, an unsupervised learning algorithm as K-means was a clear path to process the information obtained. A hierarchical clustering would be inefficient and extracting insights from it would be difficult due to the high number of categories of venues added to the data analysis.

The variables the clustering algorithm used were the ratio of families and the number of places by category. There was only one important input to define which was the number of clusters to ask K-means to build.

A small number of clusters would produce too many different suburbs in a single group which is not convenient for the allocation of resources by the local government. Likewise, too many groups would make difficult the task of identifying patterns and distribute the necessary resources and team of experts to each group.

A method that it is usually used for this purpose is the Elbow method which helps to identify the optimum number of clusters based on the distortion. This is the sum of squared distances from each point to its assigned centroid. (*Figure 1*) shows a graph plotting the distortion vs the number of clusters.

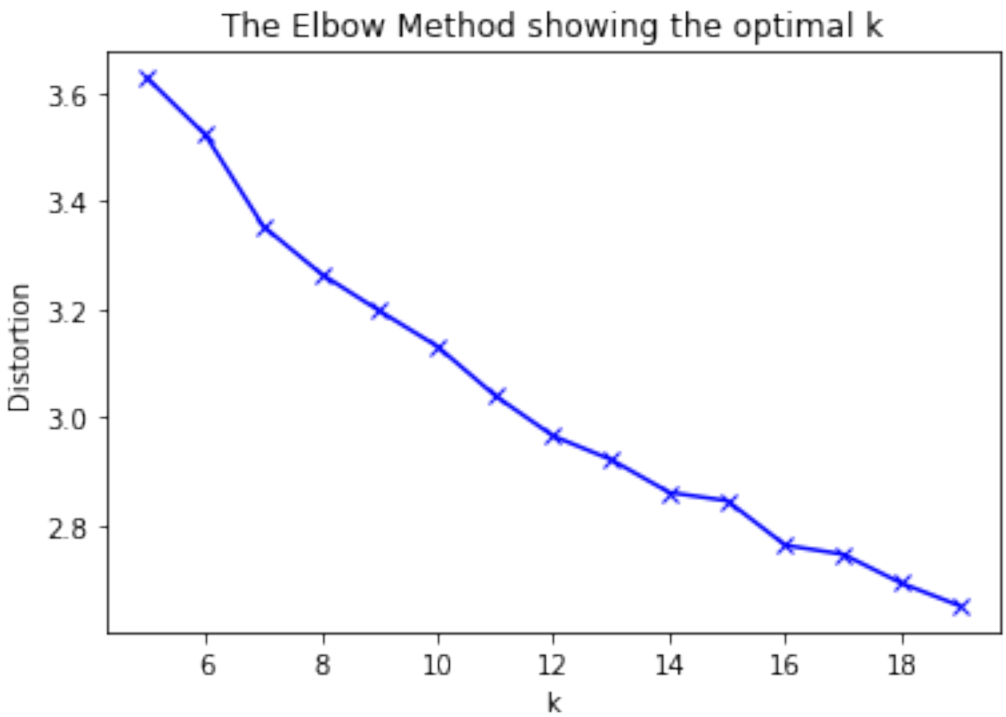

Figure 1. Plotting the Elbow Method.

Since it was not clear what the optimum point was, another version of the method was used that helped to identify the best k parameter automatically. (*Figure 2*) shows a graph using this method with 13 being the optimum k value. This tool is detailed in the following link https://www.scikit-yb.org/en/latest/api/cluster/elbow.html

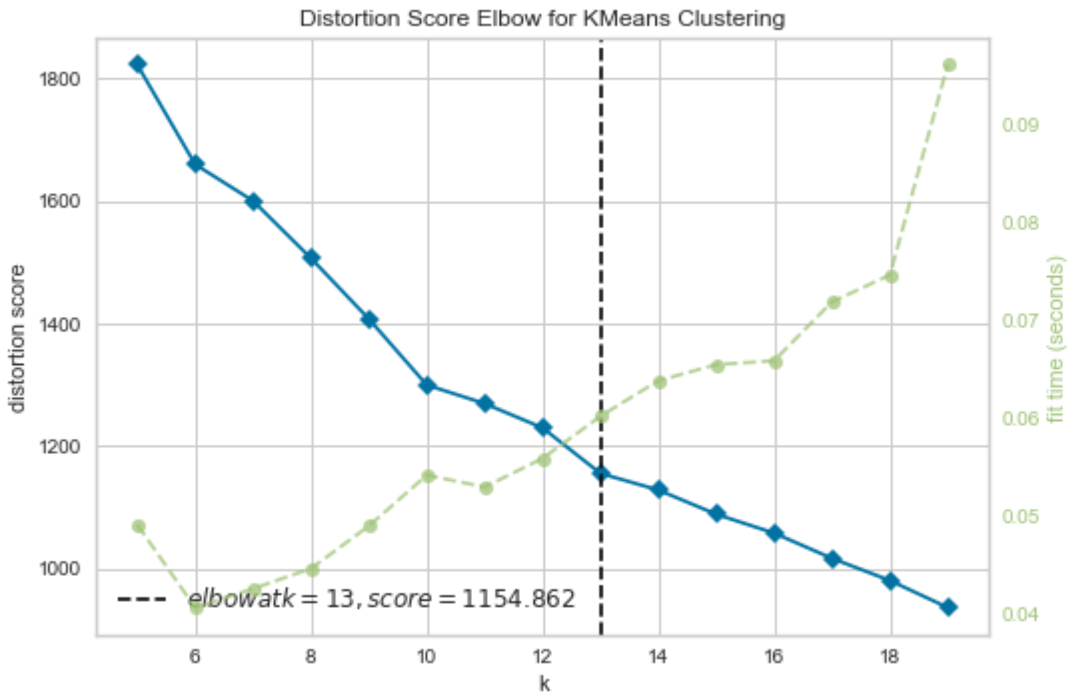

Figure 1. Plotting the Elbow Method using Yellowbrick: Machine Learning Visualization

As mentioned in the previous section, one of the requirements for the study was to identify which variables to use to feed the algorithm to produce insight. Discussions with the local government were held to know which group of the population needed to be targeted. Then, based on the final goal, is was necessary to define the categories of venues suitable for families with children.

We all the right questions to solve, the next step was to select the sources of information and tools to produce a solution. The Australian Bureau of Statistics website and datasets from the local government were selected as the demographics source of information, while the Foursquare API was the perfect fit to look up for places suitable for the study.

After fitting the data with the K-means algorithm and 13 possible clusters the suburbs were ready to be represented on a map in different colours identifying each cluster. (*Figure 3*) shows the final result.

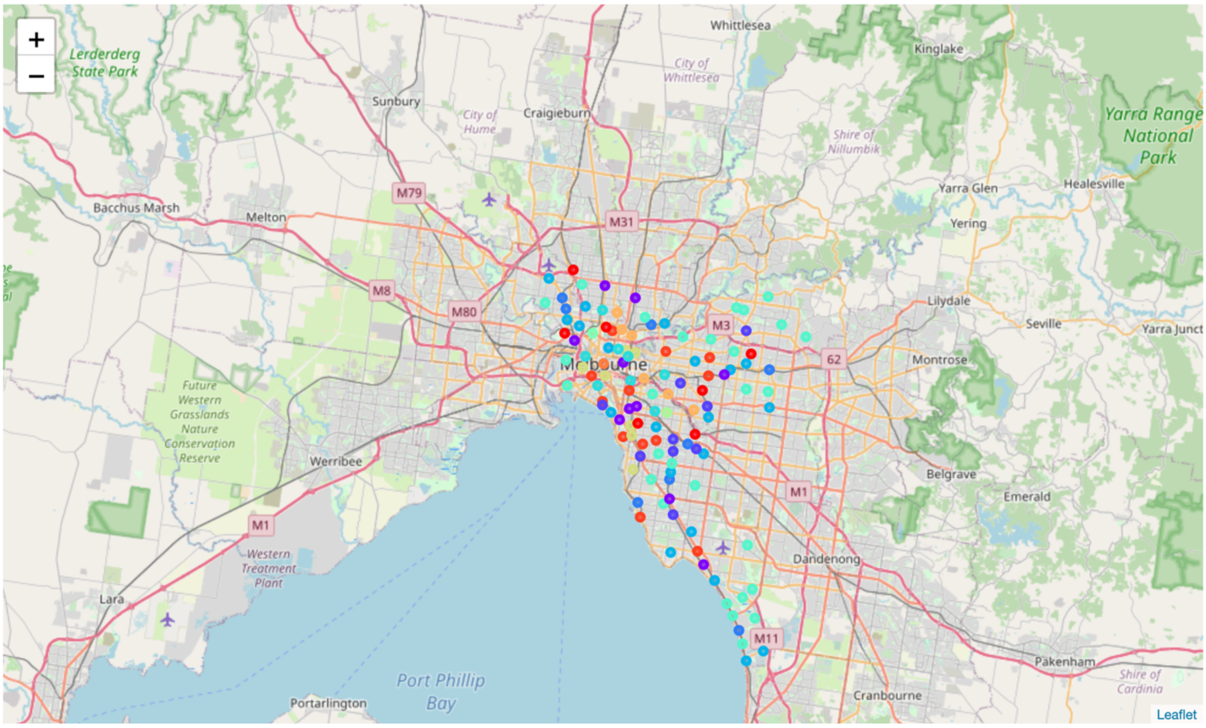

Figure 3. Clusters of suburbs for `Melbourne – Inner`

Results

The following are the 13 clusters built with the K-means algorithm. Each cluster's data frame was transposed to make it easier to read. So, the column **count** shows the number of suburbs in the cluster. The column **mean** shows the average of the corresponding venue category among the suburbs in the cluster.

The categories with no venues were removed from the clusters.

Cluster 0:

	count	mean
Family ratio	7.0	0.372318
Bowling Green	7.0	0.142857
Breakfast Spot	7.0	0.285714
Café	7.0	2.571429
Coffee Shop	7.0	0.285714
Dog Run	7.0	0.285714
Garden	7.0	0.142857
General Entertainment	7.0	1.285714

	count	mean
Historic Site	7.0	0.428571
Park	7.0	3.428571
Pizza Place	7.0	1.285714
Playground	7.0	0.857143
Plaza	7.0	0.142857
Public Art	7.0	0.142857
Sports Club	7.0	0.428571
Zoo Exhibit	7.0	0.285714

Cluster 1:

	count	mean
Family ratio	10.0	0.315437
American Restaurant	10.0	0.100000
Art Gallery	10.0	1.000000
Art Studio	10.0	0.100000
Australian Restaurant	10.0	0.100000
Bakery	10.0	0.100000
Bar	10.0	0.100000
Bridge	10.0	0.200000
Cafeteria	10.0	0.100000
Café	10.0	11.600000
Coffee Shop	10.0	1.700000
Deli / Bodega	10.0	0.200000
Dessert Shop	10.0	0.100000
Diner	10.0	0.100000
Dog Run	10.0	0.100000
Exhibit	10.0	0.100000

	count	mean
Fast Food Restaurant	10.0	0.100000
Food Court	10.0	0.100000
Football Stadium	10.0	0.100000
French Restaurant	10.0	0.100000
Furniture / Home Store	10.0	0.100000
Garden	10.0	0.300000
Garden Center	10.0	0.100000
Gastropub	10.0	0.100000
General Entertainment	10.0	0.400000
Grocery Store	10.0	0.100000
Gym	10.0	0.100000
Historic Site	10.0	0.200000
History Museum	10.0	0.100000
Indie Movie Theater	10.0	0.100000
Italian Restaurant	10.0	0.400000
Multiplex	10.0	0.100000
Park	10.0	1.900000
Performing Arts Venue	10.0	0.100000
Pizza Place	10.0	2.700000
Planetarium	10.0	0.100000
Playground	10.0	1.300000
Plaza	10.0	0.100000
Pool	10.0	0.100000
Pool Hall	10.0	0.100000
Recreation Center	10.0	0.100000
River	10.0	0.100000

	count	mean
Sculpture Garden	10.0	0.100000
Soccer Field	10.0	0.400000
Sports Club	10.0	0.100000
Supermarket	10.0	0.100000
Thai Restaurant	10.0	0.100000
Theater	10.0	0.100000
Theme Park	10.0	0.200000
Trail	10.0	0.100000
Vegetarian / Vegan Restaurant	10.0	0.100000
Zoo Exhibit	10.0	0.200000

Cluster 2:

	count	mean
Family ratio	9.0	0.361478
Art Gallery	9.0	0.666667
Bakery	9.0	0.111111
Beach	9.0	0.111111
Bike Shop	9.0	0.111111
Breakfast Spot	9.0	0.222222
Cafeteria	9.0	0.111111
Café	9.0	8.111111
Coffee Shop	9.0	1.222222
Dessert Shop	9.0	0.111111
Dog Run	9.0	0.111111
Food Court	9.0	0.222222
Football Stadium	9.0	0.111111
Garden	9.0	0.222222

	count	mean
General Entertainment	9.0	1.000000
Indie Movie Theater	9.0	0.222222
Italian Restaurant	9.0	0.333333
Movie Theater	9.0	0.555556
Multiplex	9.0	0.222222
Park	9.0	0.888889
Pizza Place	9.0	0.666667
Playground	9.0	1.666667
Plaza	9.0	0.111111
Polish Restaurant	9.0	0.111111
Public Art	9.0	0.111111
Recreation Center	9.0	0.111111
Soccer Field	9.0	0.222222
Sports Club	9.0	0.111111
Trail	9.0	0.111111
Water Park	9.0	0.111111

Cluster 3:

	count	mean
Family ratio	8.0	0.380283
American Restaurant	8.0	0.125000
Art Gallery	8.0	0.375000
Beach	8.0	0.500000
Bowling Green	8.0	0.250000
Café	8.0	18.625000
Car Wash	8.0	0.125000
City Hall	8.0	0.125000

	count	mean
Coffee Shop	8.0	1.500000
Cricket Ground	8.0	0.125000
Deli / Bodega	8.0	0.125000
Dog Run	8.0	0.125000
Eastern European Restaurant	8.0	0.125000
Food & Drink Shop	8.0	0.125000
Football Stadium	8.0	0.125000
Garden	8.0	0.250000
General Entertainment	8.0	0.750000
Indie Movie Theater	8.0	0.125000
Italian Restaurant	8.0	0.125000
Kebab Restaurant	8.0	0.125000
Middle Eastern Restaurant	8.0	0.125000
Park	8.0	1.000000
Pizza Place	8.0	2.625000
Playground	8.0	0.625000
Plaza	8.0	0.125000
Pool Hall	8.0	0.125000
Recreation Center	8.0	0.250000
Soccer Field	8.0	0.125000
Sports Club	8.0	0.125000

Cluster 4:

	count	mean
Family ratio	20.0	0.352785
Art Gallery	20.0	0.100000

	count	mean
Australian Restaurant	20.0	0.050000
Bakery	20.0	0.050000
Bar	20.0	0.100000
Beach	20.0	0.300000
Breakfast Spot	20.0	0.050000
Bridge	20.0	0.100000
Burger Joint	20.0	0.050000
Café	20.0	5.250000
Car Wash	20.0	0.050000
Cheese Shop	20.0	0.050000
Coffee Shop	20.0	1.050000
Dessert Shop	20.0	0.150000
Dog Run	20.0	0.200000
Exhibit	20.0	0.100000
Falafel Restaurant	20.0	0.050000
Fish & Chips Shop	20.0	0.050000
Flower Shop	20.0	0.050000
Food Court	20.0	0.100000
Garden	20.0	0.100000
General Entertainment	20.0	0.300000

	count	mean
Greek Restaurant	20.0	0.050000
Historic Site	20.0	0.150000
History Museum	20.0	0.050000
Indie Movie Theater	20.0	0.050000
Indoor Play Area	20.0	0.050000
Italian Restaurant	20.0	0.300000
Memorial Site	20.0	0.100000
Monument / Landmark	20.0	0.050000
Movie Theater	20.0	0.050000
Museum	20.0	0.100000
Nature Preserve	20.0	0.050000
Outdoor Event Space	20.0	0.050000
Park	20.0	1.500000
Pier	20.0	0.050000
Pizza Place	20.0	2.250000
Playground	20.0	0.500000
Plaza	20.0	0.050000
Pool	20.0	0.050000
Pool Hall	20.0	0.100000
Recreation Center	20.0	0.150000

	count	mean
Restaurant	20.0	0.050000
Soccer Field	20.0	0.100000
Sports Club	20.0	0.150000
Surf Spot	20.0	0.050000
Tattoo Parlor	20.0	0.050000
Trail	20.0	0.100000
Vegetarian / Vegan Restaurant	20.0	0.050000

Cluster 5:

	count	mean
Family ratio	8.0	0.289903
Art Gallery	8.0	2.750000
Art Museum	8.0	0.125000
Bakery	8.0	0.250000
Bar	8.0	0.125000
Breakfast Spot	8.0	0.500000
Bridge	8.0	0.625000
Bubble Tea Shop	8.0	0.125000
Café	8.0	16.125000
Coffee Shop	8.0	2.750000
Concert Hall	8.0	0.125000
Cupcake Shop	8.0	0.125000
Dog Run	8.0	0.125000
Exhibit	8.0	0.125000

	count	mean
Food Court	8.0	0.125000
French Restaurant	8.0	0.125000
Garden	8.0	0.125000
General Entertainment	8.0	0.250000
Grocery Store	8.0	0.250000
Hookah Bar	8.0	0.125000
Italian Restaurant	8.0	0.125000
Park	8.0	0.625000
Pizza Place	8.0	1.125000
Playground	8.0	0.250000
Plaza	8.0	0.125000
Pool	8.0	0.125000
Pub	8.0	0.250000
Recreation Center	8.0	0.125000
Strip Club	8.0	0.125000
Theme Park	8.0	0.250000
Vegetarian / Vegan Restaurant	8.0	0.125000
Wine Bar	8.0	0.125000

Cluster 6:

	count	mean
Family ratio	28.0	0.374008
Athletics & Sports	28.0	0.035714
Bakery	28.0	0.071429
Beach	28.0	0.321429
Café	28.0	1.392857
Coffee Shop	28.0	0.250000

	count	mean
Deli / Bodega	28.0	0.035714
Dessert Shop	28.0	0.035714
Dog Run	28.0	0.107143
Food Court	28.0	0.035714
Garden	28.0	0.107143
General Entertainment	28.0	0.178571
Italian Restaurant	28.0	0.035714
Office	28.0	0.035714
Other Great Outdoors	28.0	0.035714
Park	28.0	0.428571
Pizza Place	28.0	0.464286
Playground	28.0	0.464286
Pool	28.0	0.071429
Recreation Center	28.0	0.178571
Soccer Field	28.0	0.142857
Sports Club	28.0	0.071429
Surf Spot	28.0	0.035714
Tennis Stadium	28.0	0.035714
Trail	28.0	0.071429

Cluster 7:

	count	mean
Family ratio	1.0	0.255165
Café	1.0	1.000000
Dog Run	1.0	1.000000
Garden	1.0	1.000000
General Entertainment	1.0	3.000000

	count	mean
Hockey Arena	1.0	1.000000
Park	1.0	4.000000
Sports Club	1.0	2.000000
Zoo	1.0	1.000000
Zoo Exhibit	1.0	16.000000

Cluster 8:

	count	mean
Family ratio	1.0	0.369052
Café	1.0	26.000000
Coffee Shop	1.0	4.000000
Cricket Ground	1.0	2.000000
Dog Run	1.0	2.000000
Garden	1.0	2.000000
Italian Restaurant	1.0	2.000000
Park	1.0	2.000000
Pizza Place	1.0	4.000000
Playground	1.0	4.000000

Cluster 9:

	count	mean
Family ratio	7.0	0.280073
Art Gallery	7.0	1.857143
Asian Restaurant	7.0	0.142857
Bakery	7.0	0.142857
Bar	7.0	0.285714
Bridge	7.0	0.285714
Cafeteria	7.0	0.142857

	count	mean
Café	7.0	12.285714
Coffee Shop	7.0	3.285714
Cultural Center	7.0	0.142857
Dog Run	7.0	0.142857
Exhibit	7.0	0.142857
Garden	7.0	0.714286
Gastropub	7.0	0.142857
General Entertainment	7.0	1.714286
Grocery Store	7.0	0.142857
Gym / Fitness Center	7.0	0.142857
Historic Site	7.0	0.142857
History Museum	7.0	0.285714
Hotel Bar	7.0	0.142857
Indie Movie Theater	7.0	0.142857
Italian Restaurant	7.0	0.142857
Movie Theater	7.0	0.428571
Park	7.0	2.142857
Performing Arts Venue	7.0	0.142857
Pet Store	7.0	0.142857
Pizza Place	7.0	1.285714
Playground	7.0	0.285714
Plaza	7.0	0.428571
Pool	7.0	0.285714
Public Art	7.0	0.142857
Recreation Center	7.0	0.142857
River	7.0	0.142857

	count	mean
Street Art	7.0	0.142857
Strip Club	7.0	0.142857
Theme Park	7.0	0.142857
Zoo Exhibit	7.0	0.142857

Cluster 10:

	count	mean
Family ratio	6.0	0.321004
Art Gallery	6.0	0.500000
Bakery	6.0	0.500000
Bar	6.0	0.333333
Bowling Green	6.0	0.166667
Breakfast Spot	6.0	0.166667
Bridge	6.0	0.333333
Cafeteria	6.0	0.166667
Café	6.0	8.166667
Coffee Shop	6.0	1.666667
Design Studio	6.0	0.166667
Dog Run	6.0	0.333333
Exhibit	6.0	0.166667
Football Stadium	6.0	0.166667
Furniture / Home Store	6.0	0.166667
Garden	6.0	0.833333
Garden Center	6.0	0.166667
General Entertainment	6.0	0.666667
Ice Cream Shop	6.0	0.166667
Japanese Restaurant	6.0	0.166667

	count	mean
Park	6.0	5.000000
Pizza Place	6.0	1.500000
Playground	6.0	1.000000
Pool	6.0	0.500000
Pool Hall	6.0	0.166667
Sandwich Place	6.0	0.166667
Soccer Field	6.0	0.500000
Sports Club	6.0	0.166667
Stadium	6.0	0.166667
Strip Club	6.0	0.166667
Tennis Stadium	6.0	0.166667
Theme Park	6.0	0.166667
Trail	6.0	0.500000
Vegetarian / Vegan Restaurant	6.0	0.166667

Cluster 11:

	count	mean
Family ratio	1.0	0.154968
Bar	1.0	1.000000
Café	1.0	3.000000
Coffee Shop	1.0	12.000000
Donut Shop	1.0	1.000000
Food Court	1.0	2.000000
General Entertainment	1.0	1.000000
Juice Bar	1.0	1.000000
Library	1.0	1.000000
Movie Theater	1.0	1.000000

	count	mean
Multiplex	1.0	1.000000
Plaza	1.0	2.000000
Pub	1.0	1.000000
Shopping Mall	1.0	3.000000

Cluster 12:

	count	mean
Family ratio	12.0	0.311555
Art Gallery	12.0	0.333333
Australian Restaurant	12.0	0.083333
Bakery	12.0	0.166667
Bar	12.0	0.083333
Beach	12.0	0.083333
Breakfast Spot	12.0	0.166667
Bridge	12.0	0.416667
Café	12.0	14.750000
Coffee Shop	12.0	1.166667
Dance Studio	12.0	0.083333
Deli / Bodega	12.0	0.166667
Dessert Shop	12.0	0.083333
Dog Run	12.0	0.166667
Fast Food Restaurant	12.0	0.083333
Food Court	12.0	0.083333

	count	mean
Football Stadium	12.0	0.083333
French Restaurant	12.0	0.083333
Garden	12.0	0.416667
General Entertainment	12.0	0.750000
Historic Site	12.0	0.166667
Indie Movie Theater	12.0	0.083333
Indoor Play Area	12.0	0.083333
Italian Restaurant	12.0	0.166667
Lake	12.0	0.166667
Park	12.0	1.583333
Pizza Place	12.0	1.583333
Playground	12.0	0.916667
Plaza	12.0	0.250000
Pool	12.0	0.250000
Pool Hall	12.0	0.166667
Radio Station	12.0	0.083333
Recreation Center	12.0	0.166667
River	12.0	0.166667
Sports Club	12.0	0.083333
Trail	12.0	0.416667
Turkish Restaurant	12.0	0.083333

Discussion

Based on the goal requested by the local government and the analysis of the results, the decision of what groups of suburbs to support can be made following these criteria:

- Suburbs with a low ratio of families with children should be of low priority (e.g. clusters 5, 7, 9, 11).
- Suburbs with a high ratio of families with children and a high number of suitable venues should be of low priority to create new facilities. The focus here should be on maintenance of the current infrastructure and businesses (e.g. 8, 10, 12).
- Suburbs with a high ratio of families with children and a low number of suitable venues should be the focus of the government plan and give top priority to these areas (e.g. 0, 1, 2, 3, 4, 6).

It is important to mention the limitations of this study and how to improve it.

- The venues gathered per suburbs were only those located in a radius of 500 meters from the centre of the suburb and limiting the search to only 30 places. A substantial improvement would be not to limit the number of venues to search. Also, making the radius big enough to go out of each suburb, then filter out those venues not located in the suburb by matching the address.
- The elbow method to select the number of clusters was not the most appropriate way to select groups of clusters in this case. There is no evident change in the curve of the graph that indicates an optimal number. This may be due to the high diversity of data among the suburbs in Melbourne that makes it difficult to identify isolated groups. A more realistic approach would be for the local government to define how many groups they want to work with, depending on budget, resources and time to tackle the field studies and projects.
- Another point to be aware of is that this study was done only for the SA4 area 'Melbourne - Inner' which is the most populated of the SA4 areas conforming Melbourne. However, the local government will be interested in replicating this study to the rest of the suburbs of the city: Melbourne - North West, Melbourne - West, Melbourne - North East, Melbourne - Inner East, Melbourne - South East, Melbourne - Inner South and Melbourne - Outer East.

It is recommended for further analysis to consider the use of DBSCAN as an alternative to K-means to identify the groups of similar suburbs. Another approach can be to replicate the study with three different number of k parameter for K-means that matches the suitable number of groups which is feasible to work with. Then, produce three reports for the local government to choose the best for the task.

Conclusion

The study presented here is a good start point for the local government of Melbourne to solve the problem of suitable venues for families with children. With a full licence of Foursquare to request data about venues, plus gathering information from government archives and other online services like Google Places, the dataset of venues would be much more valuable for a better outcome.

It is critical to have discussions with stakeholders and contractors in charge of the project and logistics to define what number of clusters is the most appropriate.

After having a plan base on this study, further discussion with local councils can be of great feedback to make it more accurate and find the best possible distribution of the local resources in the area.