Finding the Best Neighbourhood in Brisbane for Families with Children

Ramon Rossi

7 April 2021

Table of Contents

Table of Figures

1. Introduction

1.1 Background

Families with children often move cities or even countries. It is often a challenge to know which suburb or neighbourhood to move to within a city. According to Bader (2019) and Schmidt (2020), important factors when choosing a neighbourhood include:

- safety
- schools offering excellent education
- children-friendly venues such as parks and playgrounds

The question is: What is the best suburb or neighbourhood for a family moving into a city?

1.2 Problem

The assumed problem is that a family wants to re-locate to **Brisbane, Australia**. They want a neighbourhood with common venues suitable for children, like parks, fitness centres and soccer fields. If possible, they want to be in an area where their children can go to a state (public/free) school which achieves better than average results. They would like the lowest crime rate possible. Which neighbourhood should they choose?

1.3 Interest

Families with children moving into Brisbane would be interested in this project. However, similar projects could be initiated for other cities to help families needing to move to those cities. The project has business value in that a website, for example, outlining recommended neighbourhoods for a particular city would potentially generate traffic (income) from families seeking to relocate there.

2. Data

2.1 Data Sources

There are 6 separate data sources that are utilised in this project:

2.1.1 For Postcode Data

A postcode dataset (CSV file) provided by Proctor (2021) contains all the postcodes for Australia along with their associated suburbs (neighbourhoods) and latitude/longitude data. Eventually, this cleaned-up dataset was used to place each neighbourhood/postcode on a map to visualise similar clusters of neighbourhoods.

2.1.2 For Crime Data

The Queensland Police Service (2021) provides an online app to view statistics (and download data in the form on an Excel file) on crimes related to a Queensland postcode for the past 2 year period. Data can be downloaded on a maximum of 10 postcodes at a time. Therefore, the Brisbane postcodes were entered 10 a time and each Excel file downloaded. The result was a total of 8 Excel files containing the crime data for all Brisbane postcodes.

2.1.3 School Data

The Queensland Government (2020) provides an CSV file with data about all Queensland schools, including the postcode and suburb a school is located in. This file does not include any performance data about the schools.

2.1.4 School Performance Data

The Queensland Curriculum and Assessment Authority (QCAA, 2019) provide a PDF file of the test results for all Queensland schools from 2019.

2.1.5 Foursquare Venue Data

Via the Foursquare API, a dataset can be generated for all venues related to the suburbs/neighbourhoods in Brisbane (450 venues total).

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	BRISBANE, BRISBANE ADELAIDE STREET, BRISBANE C...	-27.468544	153.022459	W Brisbane	-27.470120	153.021970	Hotel
1	BRISBANE, BRISBANE ADELAIDE STREET, BRISBANE C...	-27.468544	153.022459	Cartel Coffee	-27.468448	153.024845	Coffee Shop
2	BRISBANE, BRISBANE ADELAIDE STREET, BRISBANE C...	-27.468544	153.022459	Museum of Brisbane	-27.469028	153.023837	History Museum
3	BRISBANE, BRISBANE ADELAIDE STREET, BRISBANE C...	-27.468544	153.022459	Coffee Iconic	-27.469269	153.021740	Café
4	BRISBANE, BRISBANE ADELAIDE STREET, BRISBANE C...	-27.468544	153.022459	Felix for Goodness	-27.469493	153.024300	Café

Figure 1 Venue Data Generated Using Foursquare API

2.1.6 Geopy Location Data

Nominatim (part of Geopy's Python library) was used to find the latitude and longitude of Brisbane. These co-ordinates were then used to zoom into Brisbane on the Folium world map.

2.2 Data Cleaning

2.2.1 For Postcode Data

Postcodes cover more than one suburb/neighbourhood, so neighbourhoods were concatenated for each postcode. The postcodes were filtered to include only ones belonging to Brisbane.

Result of data cleaning: A data frame (see Figure 2) of all the postcodes in Brisbane (74 rows), along with their suburbs/neighbourhoods and latitude and longitude.

	Postcode	Area	Neighbourhood	Longitude	Latitude
0	4000	Brisbane Inner City	BRISBANE, BRISBANE ADELAIDE STREET, BRISBANE C...	153.022459	-27.468544
1	4005	Brisbane Inner City	NEW FARM, TENERIFFE	153.046752	-27.463097
2	4006	Brisbane Inner City	BOWEN BRIDGE, BOWEN HILLS, BRISBANE EXHIBITION...	153.175242	-27.366180
3	4007	Brisbane Inner City	ASCOT, BRISBANE AIRPORT, DOOMBEN, HAMILTON, HA...	153.061914	-27.436088
4	4008	Brisbane - North	BRISBANE AIRPORT, BULWER ISLAND, MEEANDAH, MYR...	153.136496	-27.397546

Figure 2 Postcode Data (sample)

2.2.2 For Crime Data

As there are 8 raw data files, each file was processed, grouped, and appended to one data frame. As each type of crime is listed by the date and postcode, the data was grouped by postcode and the number of crimes for each postcode was counted.

Result of data cleaning: A data frame (see Figure 3) of Brisbane postcodes along with how many crimes were committed in each postcode over the last 2 years.

	Postcode	Acts of Crime
0	4000	21928
1	4005	2396
2	4006	15145
3	4007	2359
4	4008	1089

Figure 3 Crime Data (sample)

2.2.3 School Data

This data was filtered to include only a) Brisbane schools and b) schools which belong to the type 'State School'.

Result of data cleaning: A data frame (see Figure 4) of Brisbane schools with relevant fields: school name, suburb/neighbourhood, postcode, type of school.

	Postcode	School	Locality	Type	School Latitude	School Longitude
0	4064	Albert Park Flexible Learning Centre	Milton	Non-State School	-27.465900	153.010813
1	4157	Alexandra Hills State High School	Capalaba	State High School	-27.523768	153.215428
2	4161	Alexandra Hills State School	Alexandra Hills	State School	-27.518141	153.221279
3	4115	Algester State School	Algester	State School	-27.615675	153.031915
4	4000	All Hallows' School	Brisbane	Non-State School	-27.460894	153.032816

Figure 4 School Data (sample)

2.2.4 School Performance Data

A dataset was extracted from the PDF file. This was no simple task, as the cells in the PDF tables of school results have a complex arrangement. Different methods were attempted to make the extraction but only one method was successful.

Method A: Tabula. A Python library (Tabula) was used to directly access the PDF and convert it into tables, but for some strange reason it would lose the data on the first school at the top of every page of the PDF file.

Method B: Conversion to Excel or CSV files. Using Adobe Acrobat, the PDF was converted to an Excel and CSV file, but both formats had severe formatting problems. Attempting to convert them into tables in Python resulted in errors.

Method C: Conversion to a HTML file. Using Adobe Acrobat, the PDF was converted to HTML file format. The HTML file was then converted into rows and columns of data in Python, but it still required complicated code (and testing) to process each element of the table and turn it into a useful data frame. One problem, for example, was that each school's data was spread over either 2 or 3 rows. Another problem was that many fields contained strange characters that needed to be removed. Also, different items of numerical data had to separated. At this point, these variables had to be calculated:

2.2.4.1 Calculation of School Size

Numerical data about attendance (how many students sat the test) needed to be totalled. This gives an indication of the size of each school.

2.2.4.2 Calculation of Percentage of Students above the National Minimum Standard (NMS)

Numerical data about percentages for different year groups and across different subjects were combined and averaged for each school so that a single percentage could be saved to the data frame.

2.2.4.3 Calculation of School Score

Numerical data about test scores for different year groups and across different subjects were combined and averaged for each school so that a single test score could be saved to the data frame.

Once turned into a useable data frame using Method C, the data was filtered to include only schools in the Brisbane area. This was accomplished by using a join with the Brisbane postcode data set (2.1.1) (which contains the postcodes/neighbourhoods for Brisbane).

Since this data set uses only the school name as an identifier for a school, a join is required on this field with the school dataset (2.1.3) above to get the postcode and school type along with the school performance data into one combined dataset. At this point, the data set was filtered for only (a) Brisbane schools of (b) type 'State Schools'.

Result of data cleaning: A data frame (see Figure 5) of all Queensland schools with their name, postcode, neighbourhood, attendance, score, percent of students above NMS.

	School	Locality	Attend	Score	Percent
0	A B Paterson College	Arundel	489	556.65	99.35
1	Abercorn State School	Abercorn	4	0.00	0.00
2	Aboriginal and Islander Independent Community ...	Acacia Ridge	67	435.05	81.50
3	Acacia Ridge State School	Acacia Ridge	87	406.30	85.30
4	Agnes Water State School	Agnes Water	68	419.40	87.90

Figure 5 School Performance Data (sample)

2.2.5 Combined School and School Performance Data

After joining the clean school data (already filtered for Brisbane postcodes) and clean school performance data (not filtered because of not having a postcode field), the data frame was as shown in Figure 6. There are 110 Brisbane state schools with performance data (12 schools with no performance data were removed, after which 75 schools were removed for not being state schools).

	Postcode	School	Locality	Type	School Latitude	School Longitude	Attend	Score
0	4161	Alexandra Hills State School	Alexandra Hills	State	-27.518141	153.221279	46.0	430.0
1	4115	Algester State School	Algester	State	-27.615675	153.031915	249.0	465.0
2	4007	Ascot State School	Ascot	State	-27.432671	153.055952	223.0	486.3
3	4034	Aspley East State School	Aspley	State	-27.362313	153.023968	248.0	464.6
4	4034	Aspley State High School	Aspley	State	-27.357169	153.024213	404.0	548.7

Figure 6 Combined School and School Performance Data (sample)

2.3 Feature Selection

Although school size was calculated at 2.2.4.1, it was decided not to include this as I could not justify how this feature was relevant to find neighbourhoods in which there are schools providing excellent education. There is not a strong correlation between school size and its performance (see Section 3.3). Therefore, this was dropped as a feature.

Although the percentage of students above the NMS was calculated at 2.2.4.2, it seems this feature would be redundant as the school's score (2.2.4.3) would be sufficient to 'rate' a school's performance.

3. Exploratory Data Analysis

3.1 Converting Numerical Crime Data to Categorical Data

Crime data for Brisbane was graphed as shown in Figure 7.

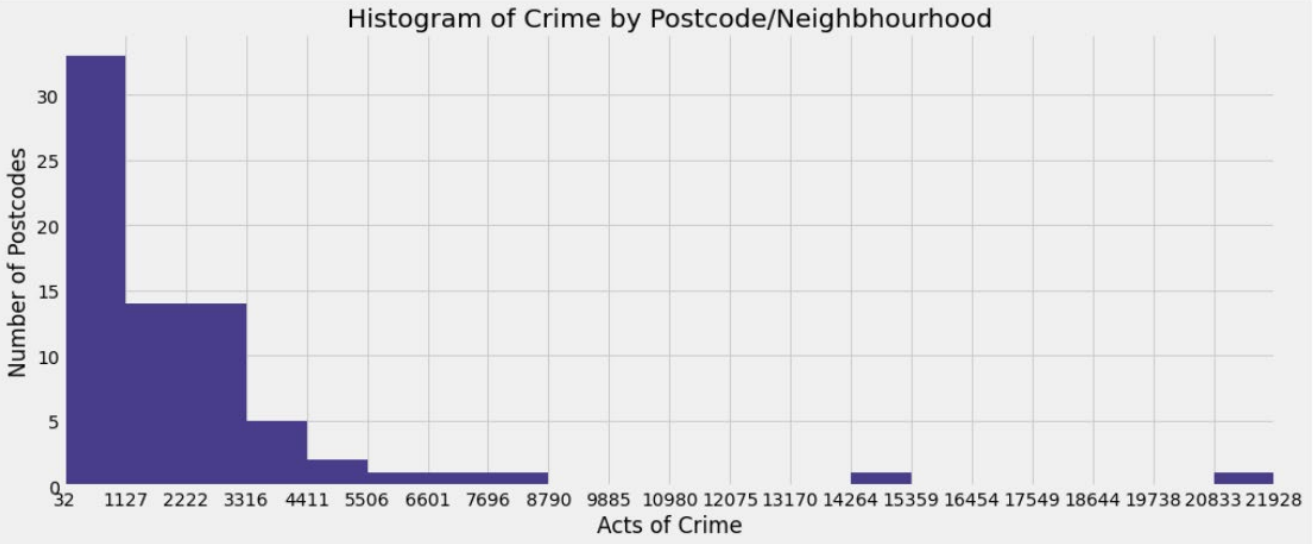

Figure 7 Histogram of Crime Data

In preparation for the k-means algorithm, the 'Acts of Crime' (numerical data in the crime dataset) was converted into categorical data according to the bins shown in Table 1.
Since 10 categories seemed adequate to differentiate the level of crime in a neighbourhood and it was a challenge to adequately invent relative labels for the crime levels, a maximum of ten bins decided upon. The two larger ranges for higher crime rates were decided because (a) the data is skewed and (b) to keep a maximum of 10 bins.

Table 1: Parameters Used to Categorise Numerical 'Acts of Crime' Field

Bin	***Range***	***Category***
1	0 to 1126.8	Extremely Low
2	1126.8 to 2221.6	Very Very Low
3	2221.6 to 3316.4	Very Low
4	3316.4 to 4411.2	Moderately Low
5	4411.2 to 5506	Fairly Low
6	5506 to 6600.8	Fairly High
7	6600.8 to 7695.6	Moderately High
8	7695.6 to 8790.4	Very High
9	8790.4 to 15359.2	Very Very High
10	15359.2 to 22000	Extremely High

3.2 Converting Numerical Score Data to Categorical Data

In preparation for the k-means algorithm, the average test score (numerical data in school performance dataset) was converted into categorical data according to the bins shown in Table 2. The histogram in Figure 8 was used to help decide on the bins.

Since 10 categories seemed adequate to differentiate to rate a school and it was a challenge to adequately invent relative labels for the school performance levels, a maximum of ten bins decided upon.

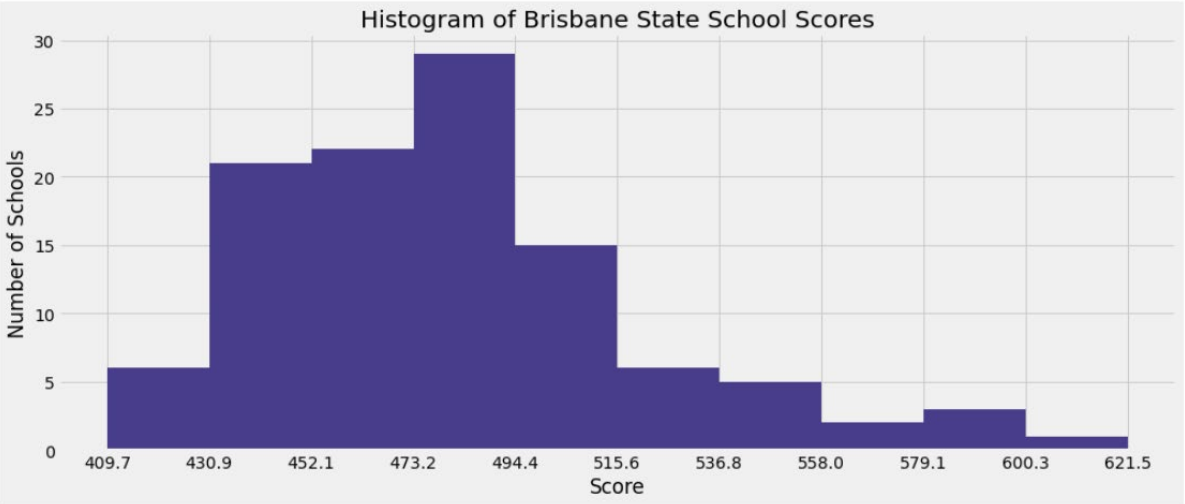

Figure 8 Histogram of School Performance Scores

The mean school score is 481.9 which approximately the middle of bin 4, therefore this bin was labelled 'Average' and the rest of the bins were labelled using this bin as a reference point.

Table 2: Parameters Used to Categorise Numerical School Performance Field

Bin	***Range***	***Category***
1	409.7 to 430.9	Significantly Below Average
2	430.9 to 452.1	Moderately Below Average
3	452.1 to 473.2	Slightly Below Average
4	473.2 to 494.4	Average
5	494.4 to 515.6	Slightly Above Average
6	515.6 to 536.8	Moderately Above Average
7	536.8 to 558	Significantly Above Average
8	558 to 579.1	Moderately Below Top Rated
9	579.1 to 600.3	Slightly Below Top Rated
10	600.3 to 621.5	Top Rated

3.3 Relationship Between School Size and Score

There is not a strong relationship (correlation score of 0.47) between a school's performance (average national test 'Score') and the school's size ('Attend' – how many attended the national test). Therefore, it was decided the Attend field was irrelevant to include as input to the k-means algorithm (see Figure 9).

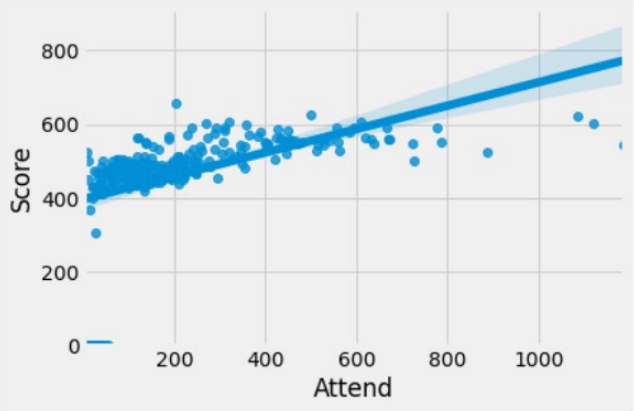

Figure 9 Chart of Correlation Between School Size and School Performance

4. Methodology

The k-means clustering algorithm was chosen to group Brisbane neighbourhood according to similar venues, crime rates and school results. To do so, the datasets for Brisbane's venues, crime rates, and school performance were one-hot encoded separately before being integrated into one dataset used as input for k-means.

4.1 Choosing the Best Value of k

To assist in choosing the best value of k, the SSE for using various values of k was graphed (See Figure 10).

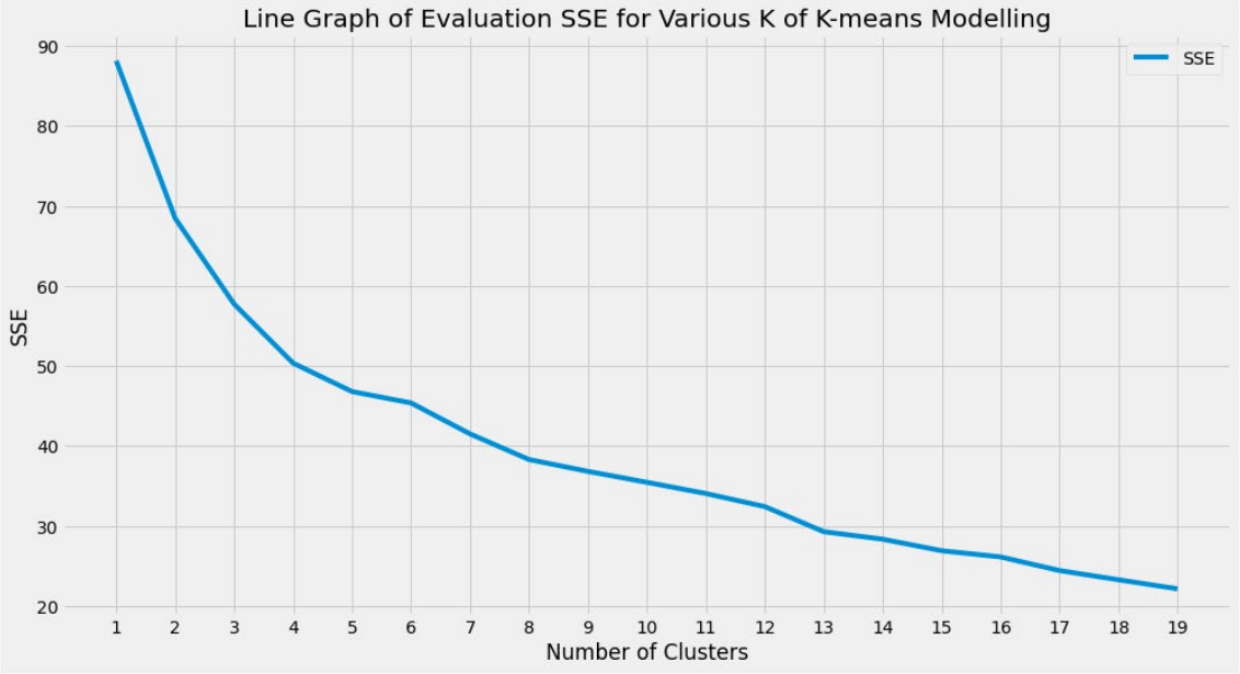

Figure 10 SSE for Various Values of k

Looking at Figure 10 and using the 'elbow method', it is difficult to decide on a clear winner for a value for k as there is no obvious elbow. However, we want to use the low SSE while finding a slight elbow. There are slight elbows at k = 4, 5, 8 and 13. I have chosen k = 8 because the next elbow at k = 13 would mean reporting on 13 clusters which is too time-consuming.

5. Results

Figure 11 shows the results of running k-means with k = 8.

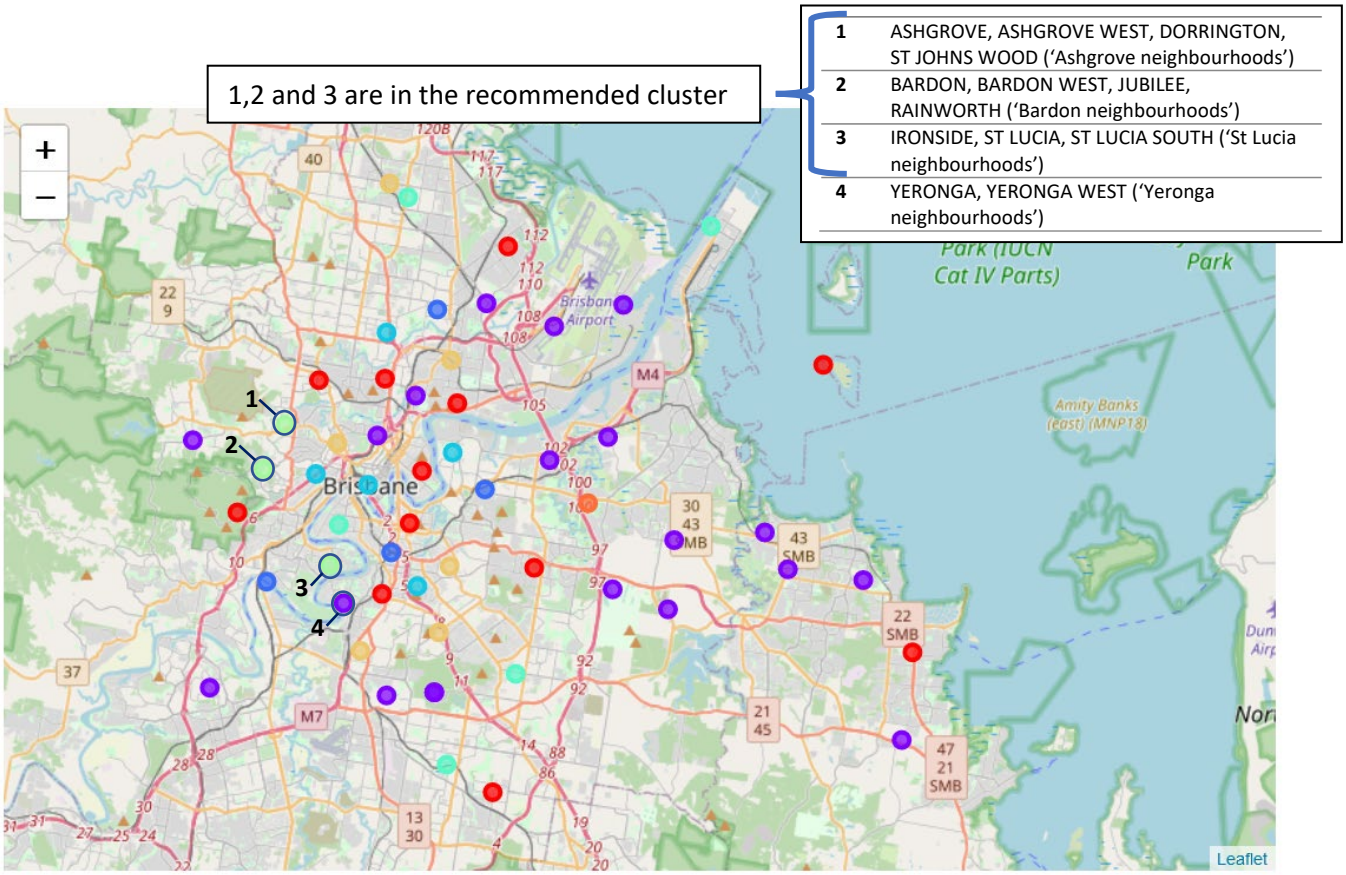

Figure 11 Results of Clustering of Neighbourhoods Using k-means

5.1 Discussion: Analysis of the Clusters Found by k-means

Only one cluster (shown in the lightest green in Figure 11) showed all the elements of the type of neighbourhood ideal for a family with children.

The neighbourhoods found in the recommended cluster most commonly have venues related to fitness, such as parks, soccer fields and gyms (see Figure 12). The most common schools perform above average (see Figure 13). All these neighbourhoods have extremely low crime levels (see Figure 14).

The neighbourhoods found in the recommended cluster are:

1. ASHGROVE, ASHGROVE WEST, DORRINGTON, ST JOHNS WOOD ('Ashgrove neighbourhoods')
2. BARDON, BARDON WEST, JUBILEE, RAINWORTH ('Bardon neighbourhoods')
3. IRONSIDE, ST LUCIA, ST LUCIA SOUTH ('St Lucia neighbourhoods')

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
22	ASHGROVE, ASHGROVE WEST, DORRINGTON, ST JOHNS ...	Ice Cream Shop	Indian Restaurant	Sports Club	Gym / Fitness Center	Chinese Restaurant
25	BARDON, BARDON WEST, JUBILEE, RAINWORTH	Soccer Field	Playground	Convenience Store	Café	Tennis Court
27	IRONSIDE, ST LUCIA, ST LUCIA SOUTH	Japanese Restaurant	Sandwich Place	Bus Stop	Indian Restaurant	Grocery Store

Figure 12 Most Common Venues in Recommended Cluster

	Neighbourhood	1st Most Common School	2nd Most Common School	3rd Most Common School
22	ASHGROVE, ASHGROVE WEST, DORRINGTON, ST JOHNS ...	School_Slightly Above Average	School_Significantly Below Average	School_Moderately Below Average
25	BARDON, BARDON WEST, JUBILEE, RAINWORTH	School_Slightly Above Average	School_Significantly Below Average	School_Moderately Below Average
27	IRONSIDE, ST LUCIA, ST LUCIA SOUTH	School_Slightly Above Average	School_Significantly Below Average	School_Moderately Below Average

Figure 13 Most Common Types of Schools in Recommended Cluster

	Neighbourhood	Crime Group
22	ASHGROVE, ASHGROVE WEST, DORRINGTON, ST JOHNS ...	Extremely Low
25	BARDON, BARDON WEST, JUBILEE, RAINWORTH	Extremely Low
27	IRONSIDE, ST LUCIA, ST LUCIA SOUTH	Extremely Low

Figure 14 Level of Crime in Recommended Cluster

5.2 Discussion: Critical Review of the Best Cluster of Neighbourhoods

Should the St Lucia neighbourhood be included in the recommended cluster?
In terms of venues, the St Lucia neighbourhood is the only one in the cluster not to have venues related to sports and fitness in its top three most common venues (see Figure 12). Should this disqualify the St Lucia neighbourhood from being recommended? According to Schmidt (2020), amenities such as 'grocery stores, convenient stores, cafes, and restaurants' are also important for families to have close by. Bader (2019) explains the importance of a grocery convenient store for parents:

> *As a parent, you want your grocery shopping to be as efficient as possible. If you forget to buy anything, you also would like the store to be close by, to make that extra run quickly, because in the next hour the kids will be back from school and then they have basketball practice.*

We can see that restaurants, eating places and grocery stores make up 3 out of the top 4 common venues for St Lucia (see Figure 12). Therefore, there is good reason to continue to recommend the St Lucia neighbourhood to families despite there not being an abundance of parks and sport facilities in that postcode. Families living in St Lucia are very close to the Yeronga neighbourhood (see Figure 11 – point 4 on the map), so can easily take advantage of the most common venue in the Yeronga neighbourhood: parks (see Figure 15).

	Neighbourhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
34	YERONGA, YERONGA WEST	6	Park	Boat Launch	Business Service

Figure 15 Most Common Venues for Yeronga Neighbourhoods

6. Conclusion

Taking into account three important factors for families with children wanting to relocate to the city of Brisbane, the k-means machine learning algorithm has been used on combined datasets involving venues, crime and school performance data to recommend one cluster of three neighbourhoods for them to relocate to. The k-means algorithm could similarly be used for other cities families may wish to relocate to by using similar datasets appropriate to those cities.

References

Web Articles

Bader, A. (2019). Best Practices for Choosing a Neighborhood for a Family [web article]. Retrieved from https://www.apartmentlist.com/renter-life/best-practices-for-choosing-a-neighborhood-for-a-family

Schmidt, D. (2020). How to Pick and Move to the Best Neighborhood for You and Your Family [web article]. Retrieved from https://www.thespruce.com/choosing-the-right-neighborhood-2435878

Data Files

QCAA. (2019). NAPLAN 2019 Outcomes – All Queensland Schools. [Data file]. Retrieved from https://www.qcaa.qld.edu.au/downloads/publications/qcaa_stats_naplan_19_outcomes.pdf

Queensland Government. (2020). State and non-state school details. [Data file]. Retrieved from http://opendata.dete.qld.gov.au/state_schools/CentreDetails_May_2020.csv

Queensland Police Service. (2021). Online Crime Map. [Data file]. Retrieved from https://qps-ocm.s3-ap-southeast-2.amazonaws.com/index.html

Proctor, M. (2021). Download Free Database of Australian Postcodes. [Data file]. Retrieved from https://www.matthewproctor.com/Content/postcodes/australian_postcodes.csv