Capstone Project

Opening a New Bakery in Brisbane, Australia

Andreas Andrianatos

January 10, 2021

Introduction

Background

Bourke Street Bakery is a bakery that started in Surry Hills, Sydney, Australia offering hand made goods, catering services as well as baking classes. It has since expanded to eleven locations within the Greater Sydney area and become quite popular with people all around Sydney.

Business Problem

The owners wish to expand their business outside of Sydney to Brisbane, Australia but are not familiar with the layout of suburbs and thus are unsure of the best location to open their first bakery in Brisbane. This project aims to examine suburbs in both Sydney and Brisbane to determine which locations would be most suitable to open a bakery.

Data Acquisition

The data required for this project is location data of each suburb in Sydney and Brisbane, venue information in both cities and details of existing bakeries. Data of venues in each suburb is also required for this project. The locations of all the data is as follows:

- Wikipedia – for a list of suburbs in Sydney and Brisbane
- Bourke St Bakery's website – for a list of existing locations
- Google's geocoding API – for latitude and longitude data
- distance.distance function from the Geopy package – to calculate distance between two coordinates
- Foursquare's API – to retrieve venue information in each suburb

Methodology

This section describes the methodology used in each step of the project.

Existing Suburbs

The locations of exiting bakeries can be found on Bourke St Bakery's website. Once the list of suburbs was found, the latitude and longitude was found using Google's geocoding API via the geocoder package in Python.

As most of the data examined in this report is geographical data, the Folium package in Python was used for visualisation purposes. In fig. 1, the locations of existing bakeries are overlayed on a map of Sydney, with each bakery represented by a hollow blue circle marker.

From the location shown in fig. 1, the majority of bakeries are in close proximity to the city, within 10km from the city centre. It would be expected that candidate suburbs in Brisbane will also be in close proximity to the centre of the city.

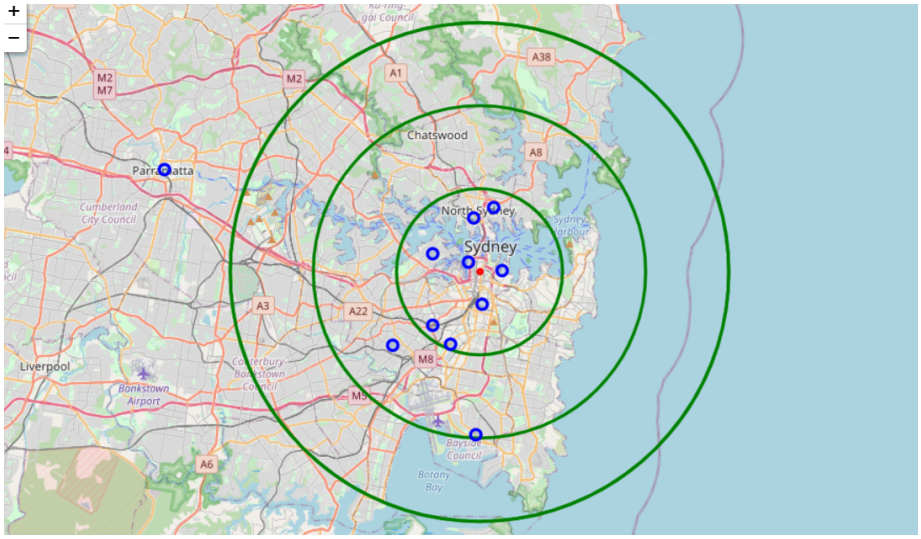

Figure 1: Locations of Bourke St Bakery in Sydney represented by hollow blue circles. The red dot locates the centre of Sydney and the green circles show 5km, 10km and 15km from the city centre.

Suburb Data for Sydney and Brisbane

Wikipedia articles were used to collect a list of suburbs for Sydney and Brisbane. The Beautiful Soup package in Python was used to parse the contents of both webpages resulting in two lists of suburbs. Once the suburbs were passed, the resulting DataFrames were checked for any duplicate suburbs, which occurred in Brisbane due the way they were categorised on Wikipedia.

The latitude and longitude of each suburb was retrieved using Google's geocoding API through the geocoder package in Python. The API allows the conversion of a text based location (i.e. an address) and returns geographical coordinates. The distance between the city centre and each suburb was calculated using **distance.distance** function from the geopy library which calculates the geodesic distance between two coordinates.

The geographic locations of each suburb in Sydney and Brisbane were plotted in fig. 2 and fig. 3 respectively. In both figures, red markers indicate suburbs that are within 15km of the city centre and to be considered in the study.

It should be noted that the sources for suburb names are defined by two different criteria, the lists being suburbs in Greater Sydney and suburbs in the City of Brisbane, which explains the large number of suburbs that are excluded from the Sydney part of the study.

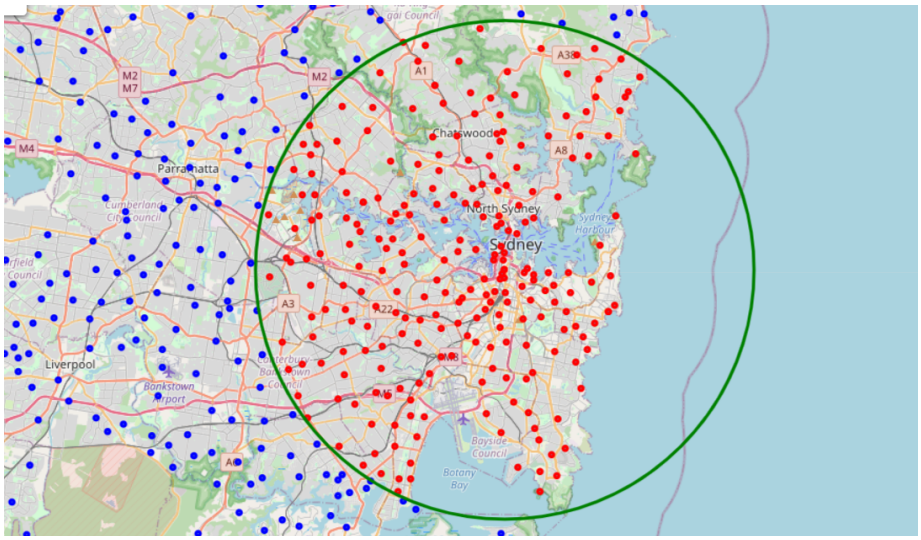

Figure 2: Suburbs within 15km of Sydney's city centre.

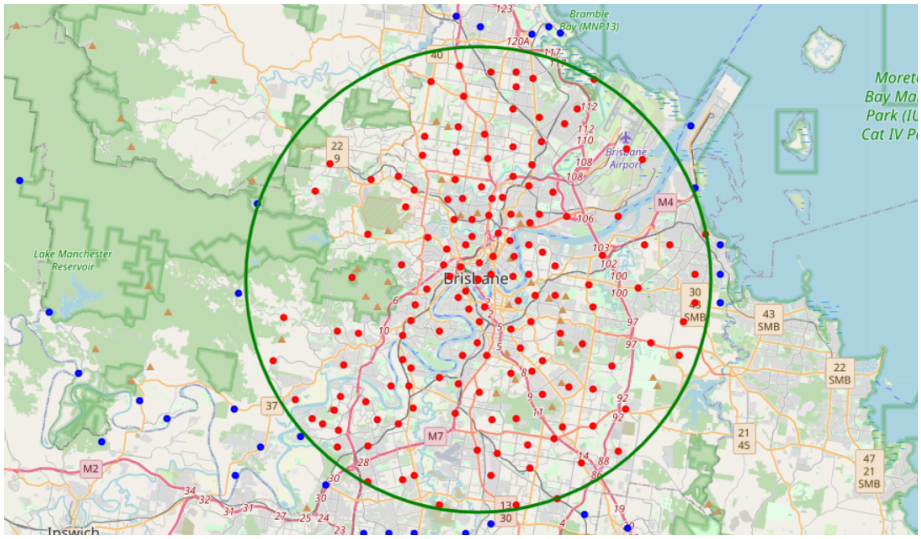

Figure 3: Suburbs within 15km of Brisbane's city centre.

Venue Data Using Foursquare

Data for venues in each suburb was obtained using the Places API provided by Foursquare. The API offers access to Foursquare's global database of venue data and user content. Venues were found using the explore reference which returns recommended venues within a radius of a specified latitude and longitude. As this study focusses on venue types, only the venue name and category are the only items of interest from the search results. The Foursquare API returns data in the json format, with the venue name and category simply extracted once the "path" to the list of items is known.

To implement the Foursquare searches in Python, a function was written that created a URL to call the API. The functions looped through each row of the DataFrames containing Sydney and Brisbane suburb information resulting in a list of venues in each city. This resulted in 12164 venues being found in Sydney and 4450 venues found in Brisbane. While using venue recommendations returns a large variety of venues compared to entering specific search terms, however, information could not be found regarding venues that do not have recommendations. It is also assumed that some venues have not been recorded on Foursquare's database but for the purpose of this study the results should still give a good representation of the venue types in each suburb.

For the case of Sydney, one of the suburbs did not return any results from Foursquare enquiries and thus will not be included in the study. All other suburbs returned at least one venue. The ten most common venue types in both cities are shown in table 1.

Sydney	**Quantity**	**Brisbane**	**Quantity**
Café	1941	Café	389
Park	476	Coffee Shop	231
Thai Restaurant	365	Supermarket	114
Coffee Shop	360	Pizza Place	109
Pub	327	Sandwich Place	93
Japanese Restaurant	300	Bakery	92
Pizza Place	300	Fast Food Restaurant	89
Bakery	294	Park	85
Bar	288	Thai Restaurant	80
Italian Restaurant	284	Liquor Store	78

Table 1: Ten most common venues in Sydney and Brisbane.

Interestingly, cafés are the most common type of venue in both cities, with Sydney having 1941 and Brisbane having 389 plus any others that have not been found using Foursquare.

Preparing the Data for Clustering

Before clustering was applied to the data, several steps were taken to prepare and sort the data. From table 1, both cities have cafés and coffee shops, these were combined to form Cafés. Similarly, pub, bar and sports bar were merged under the category of pub (note that Brisbane does not have any venues under the category of sports bar). These were the only venues merged as they were the most straightforward, possible improvements could

Sydney	Quantity	Brisbane	Quantity
Café	2301	Café	620
Pub	660	Pub	128
Park	476	Supermarket	114
Thai Restaurant	365	Pizza Place	109
Japanese Restaurant	300	Sandwich Place	93
Pizza Place	300	Bakery	92
Bakery	294	Fast Food Restaurant	89
Italian Restaurant	284	Park	85
Chinese Restaurant	182	Thai Restaurant	80
Supermarket	176	Liquor Store	78

Table 2: Ten most common venues in Sydney and Brisbane.

be made by categorising all the venues that fall under the 'Asian Restaurant' venue type. After this step was completed, the most common venues in each city is shown in table 2.

Next, to prepare the venue numbers for clustering. After applying one hot encoding, the occurrence frequency of each venue type in each suburb was found. This was done by scaling the number of each venue type by the total number of venues in each suburb, for example if a suburb has 10 venues and 2 are of Type A, Type A has a frequency occurrence of $2 \div 10 = 0.2$.

This results in values of 0–1 for each venue type in a particular suburb whereas the distance from the city centre still ranged from 0–15. The distance data was transformed using the **StandardScalar** class from Scikit-Learn. The transformation results in data with a zero mean and unit variance.

The final step was to make sure both sets of venue data had the same number of columns and same column names. Making use of sets in Python, the columns of DataFrames containing venue data were made to be the same and an occurrence frequency of 0 was given to all suburbs in the new columns (i.e. zero venues of that type in each suburb).

K-Means Clustering

K-Means clustering is an unsupervised machine learning algorithm that groups data in k clusters. Initially, the model was run considering the scaled distance between the suburb and city centre. To identify the best value of k, Sydney suburb data was fitted to the algorithm multiple times and the optimum number of clusters to use was identified using the elbow method. The KMeans class from the Scikit-Learn library was used to run the K-means algorithm and the inertia_ attribute was used to determine the distorion score. The resulting plot is shown in fig. 4.

The optimum number of clusters can be identified by the location of the "elbow" in plot and it is evident that either four or five clusters should be used. For this study five clusters were used when clustering the Sydney suburb data. The suburbs grouped in their clusters for Sydney and Brisbane are shown in fig. 5 and fig. 6 respectively.

It appears the suburbs have been grouped by distance to the city centre and it is unknown how the venues have influenced the clustering. To investigate, the clustering was repeated without considering the distance to the city centre. The results of the new clustering are shown in fig. 7 and fig. 8.

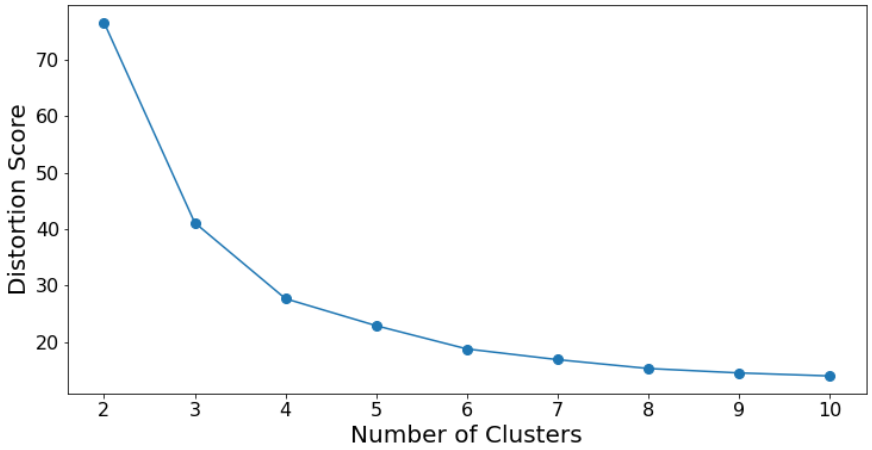

Figure 4: Suburbs within 15km of Brisbane's city centre.

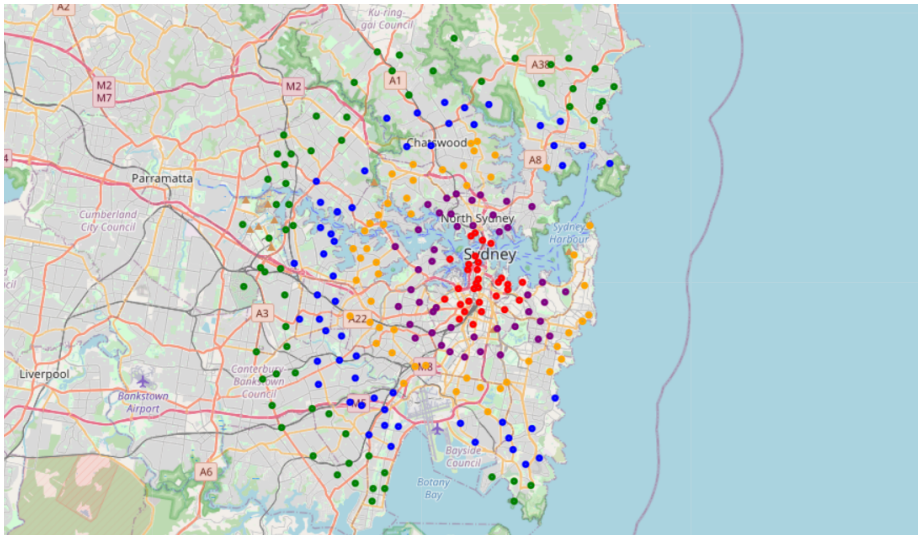

Figure 5: Suburbs of Sydney grouped in their clusters when distance to the city centre is considered.

From the new clusters it can be seen that there is no longer a dependence on distance from the city centre, with clusters appearing much more randomly geographically. In fig. 7 it is evident that clusters 2 (green), 1 (blue) and 0 (red) are the three most populous clusters with clusters 3 (purple) and 4 (yellow) accounting for only 10 of the suburbs. When examining the suburbs of Brisbane in fig. 8, almost all the suburbs belong to clusters 1 and 2.

To identify suburbs where a bakery should be opened in Brisbane, the clusters that suburbs with an existing bakery were first identified. It was found that eight bakeries belonged to cluster 2. Suburbs in Brisbane belonging to cluster 2 have been nominated as preferred suburbs to open a new bakery. These suburbs can be viewed in fig. 9.

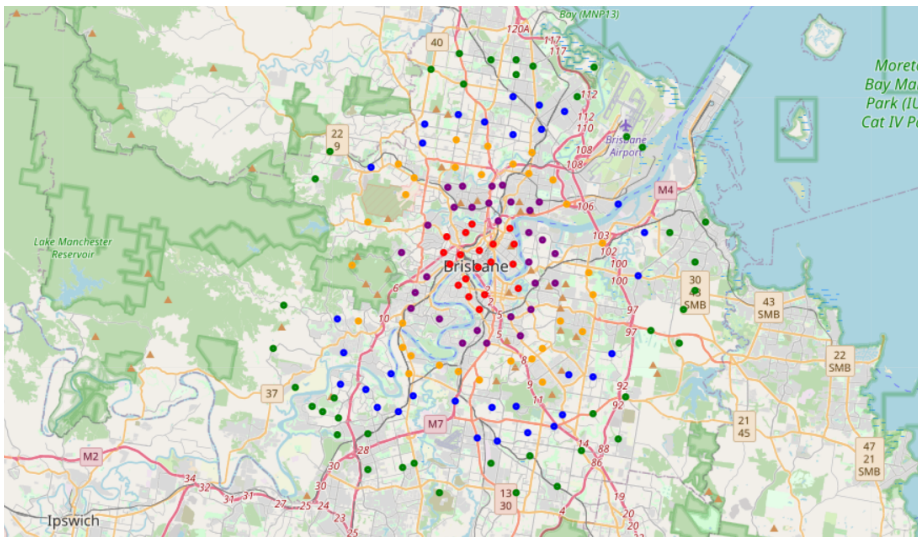

Figure 6: Suburbs of Brisbane grouped in their predicted clusters based on the Sydney data when distance to the city centre is considered.

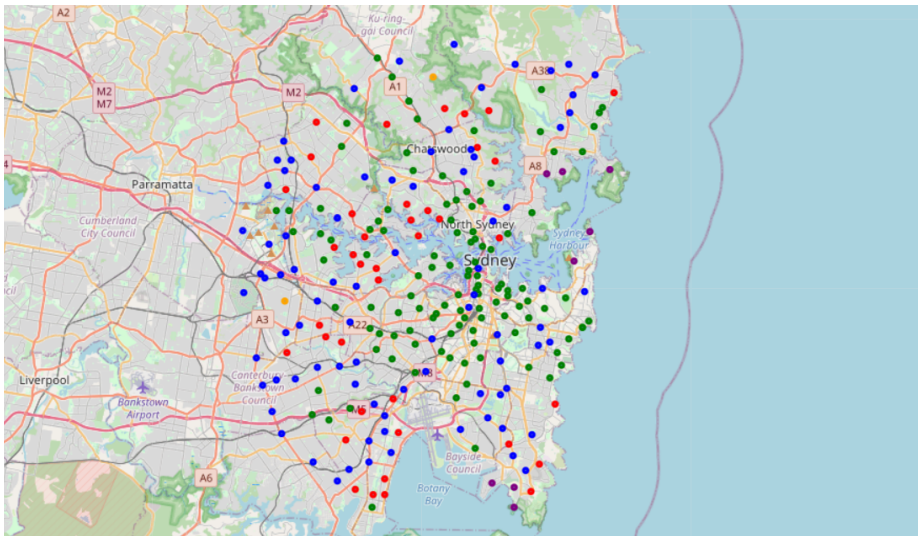

Figure 7: Suburbs of Sydney grouped in their clusters when only the venue data is considered.

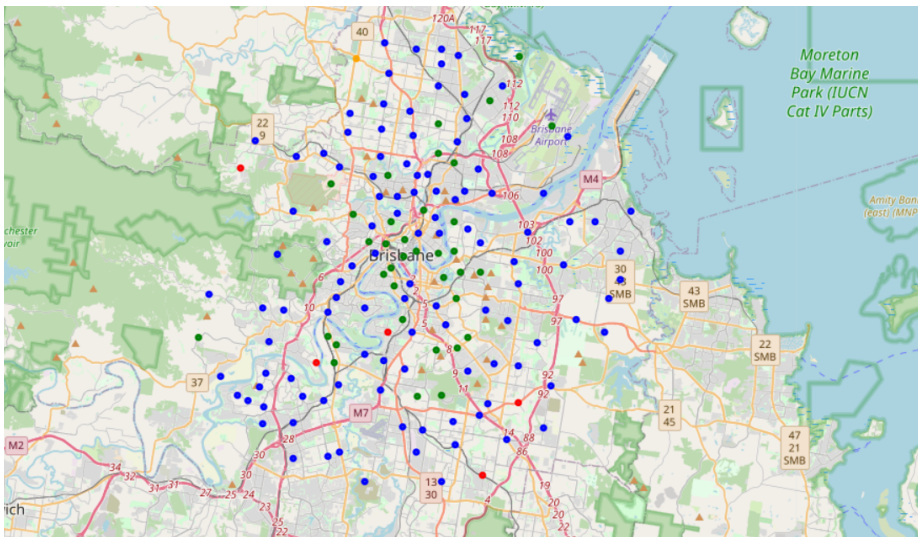

Figure 8: Suburbs of Brisbane grouped in their predicted clusters based on the Sydney data when only venues data is considered.

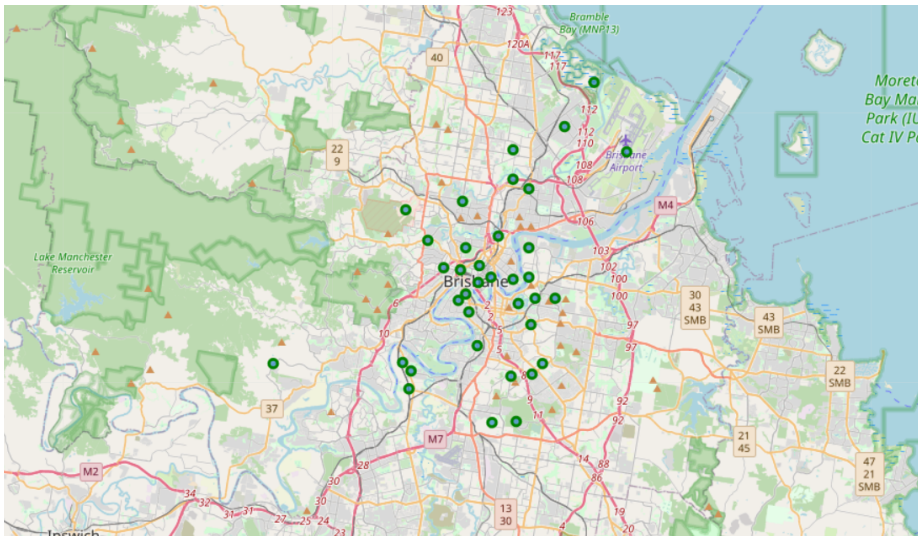

Figure 9: Suburbs of Brisbane grouped in their predicted clusters based on the Sydney data when only venues data is considered.

It is evident from fig. 9 that most of these suburbs are close to the city centre. If the results from the clusters where distance to the city centre included are considered, an intersection of the recommended suburbs can be found resulting in a more refined selection of suburbs. The results of refining the suburb list can be viewed in fig. 10 in the next section.

Results

Figure 10 shows the final list of candidate suburbs once the venue data and distance to the city centre are considered. It is evident that all the suburbs are close to the city centre. From looking at the most common venues, they are suburbs where cafés make up approximately 20% or more of the venue with the remaining venues being restaurants of various types and bars.

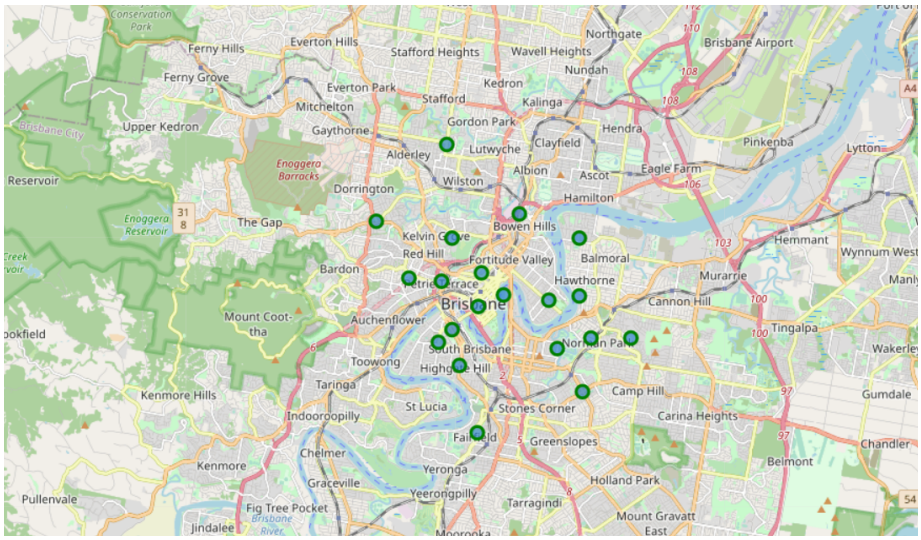

Figure 10: Candidate suburbs for a new bakery once the suburb list is was refined.

Discussion

The first step of this analysis was to find the coordinates of suburbs in both Sydney and Brisbane before searching for venues in each suburb. It was noted earlier that when searching for suburbs, a single coordinate pair was used for a location and venues were searched within a radius up to a limit of 100 items. The first potential issue is that not all suburbs are a regular shape, and there is potential that not all coordinate pairs do not locate the centre of each suburb. This may results in venues being omitted from the study or venues being included twice in two different suburbs. Capturing more venues could be improved by iterating through the area in a more ordered fashion rather than relying on suburb details.

The first attempt at clustering included a value representing distance from the city centre. It was found that the value for distance dominated the clustering algorithm and it

appeared that the suburbs were grouped based on proximity to the city centre, however, results found that eight of the ten existing bakeries belonged to the two clusters closest to the city centre which led to the distance property also being considered in the final results.

This also leads to the question, "What other features should be considered to find an ideal opening location?". This study has only considered distance to the city centre and frequency occurrence of venues in suburbs across Sydney and Brisbane. It does not consider things such as population or population density, public transport options, workplaces that may not get captured as a "venue", demographics, or average incomes, all things that could potentially influence the revenue of a bakery. While not within the scope of this study, it would be interesting to explore the effects these features would have on the results of the study.

Regardless of the aforementioned limitations, the final results of this study presents a good first view of suburbs where Bourke St Bakery could open a new store in Brisbane.

Conclusion

Bourke St Bakery, a popular bakery chain in Sydney, Australia, want to open their first store in Brisbane, Australia. This study has examined venue data in both Sydney and Brisbane to identify candidate suburbs for the opening of their first store. The candidate suburbs are:

Ashgrove, Bowen Hills, Brisbane CBD, Bulimba, Coorparoo, East Brisbane, Fairfield, Grange, Hawthorne, Highgate Hill, Kangaroo Point, Kelvin Grove, New Farm, Norman Park, Paddington, Petrie Terrace, Seven Hills, South Brisbane, Spring Hill, West End

All these suburbs are already popular locations for cafés, restaurants and bars. It is recommended that the client examine the merits of each of these locations, investigating factors which would affect the running of business and not covered in this report.