Identifying Potential Cities for Expansions

Aidan Donoghue – Coursera Capstone Project

1. Introduction

1.1. Background

A common issue for businesses is deciding on new locations to offer their services. How people behave and the commonalities of running a business in different cities.

Identifying cities where people will frequent the new business and where similar businesses have also been successful will enable business owners more insight when choosing expansion targets.

1.2. Problem

A client who runs a successful business, has asked for us to identify cities where their business could be equally successful. In this project we will try to find similar cities to a selected city ('Auckland' in New Zealand) - which is where the client's business is located.

To identify possible cities, we will base this decision based on the correlation between the City Quality of Life Index and compare similar cities based on location data from FourSquare. The quality-of-life index will give a glimpse into the general sentiment of our client's customers, while the FourSquare location data will provide insight into the typical city venue make-up.

2. Data Acquisition and Cleaning

2.1. Data Sources

For the quality-of-life dataset we have used a freely available dataset from Kaggle (City Quality of Life Dataset | Kaggle).

Additionally, FourSquare data will be used to identify the top locations identified using the quality-of-life dataset. FourSquare is able to return the surrounding venues in the city centre.

2.2. Data Cleaning

The quality-of-life dataset contained additional columns (quality indices) that were not overly relevant to a business application, therefore these were dropped from the database.

Furthermore, we applied standardisation to even out the spread of data so that it is easier to use during machine learning.

The FourSquare data does not require cleaning as it is returned based on an API request.

2.3. Data Selection

The quality of life is identified by several unique scores (housing, cost of living, startups, venture capital, travel connectivity, commute, business freedom, safety, healthcare, education, environmental quality, economy, taxation and internet access). By taking an average of the following scores:

- Cost of living
- Startups
- Travel Connectivity
- Business Freedom
- Economy
- Taxation

3. Data Acquisition and Cleaning

3.1. Calculation of Target Variable

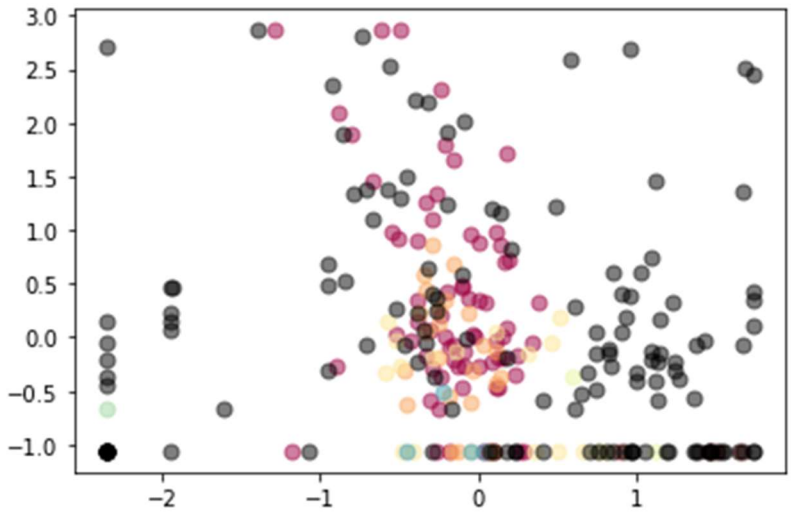
Initially we investigated a simple mean-square-error difference between every city in the quality-of-life dataset and our selected city (Auckland). This produced some reasonable results however there was no consideration of which values were close and which had large differences.

Therefore, this approach was abandoned for more advanced methods.

3.2. Data Relationship – Quality-of-life Indices

To identify relationships, we have used DBSCAN unsupervised clustering. Shown below is a scatter plot showing every city with a comparison of Cost of Living and Venture Capital.

As shown below, this particular combination of variables does not show a strong correlation together.



Plot of DBSCAN results comparing Cost of Living and Venture Capital

The results of this quality-of-life have provided a list of 70 other cities that have similar indices. To consider the most similar mean square error, this resulted in only minor changes to the order.

3.3. Data Relationship – FourSquare Data

From the similar cities database, we determined the city centre Latitude and Longitude, which was then used as the city centre location.

We have considered a 500m radius around the calculated city centre, and have not restricted the returned venues (i.e. the venues will include food shops, entertainment, or other shops).

To utilise the FourSquare data we have modified the foursquare output with one-hot encoding, which is used to help bound the classification of the cities by bounding the potential responses.

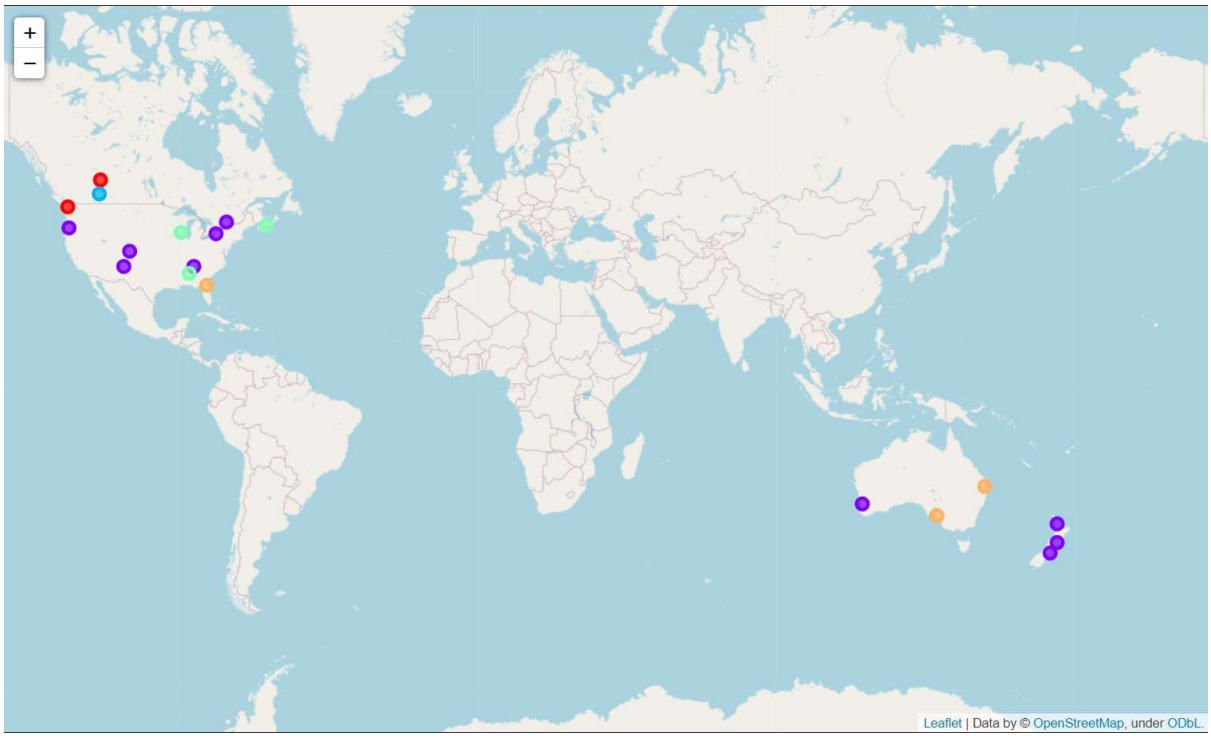
4. Machine Learning Clustering

We used two types of un-supervised machine learning techniques: DBSCAN and k-means.

DBSCAN was chosen due to the possibility that the cities form unique shaped clusters based on the various indices used to determine quality-of-life.

K-means has been utilised to segment the cities based on the surrounding venues. This was chosen due to the fairly limited nature of the types of outputs available.

The K-means clustering produced the following graphic (note: purple = points on interest):



5. Conclusion

In this study we have pulled together various cities to identify similar quality-of-life scores. Then we used venue data from FourSquare which we processed further to get a count of each type of venue within 500m radius of the city centre.

The final output was a selection of cities see below:

	UA_Name	UA_Country	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
12	Auckland	New Zealand	Theater	Japanese Restaurant	Steakhouse
260	Wellington	New Zealand	Café	Coffee Shop	Bar
185	Perth	Australia	Coffee Shop	Sushi Restaurant	Korean Restaurant
67	Colorado Springs	Colorado	Coffee Shop	Hotel	Mexican Restaurant
57	Chattanooga	Tennessee	Bar	Coffee Shop	Music Venue
47	Buffalo	New York	Hotel	Coffee Shop	Pizza Place
2	Albuquerque	New Mexico	Bar	Coffee Shop	Brewery
62	Christchurch	New Zealand	Bar	Hotel	Plaza
86	Eugene	Oregon	Coffee Shop	Concert Hall	Pizza Place
30	Birmingham	Alabama	Bar	American Restaurant	Asian Restaurant

Based on the results above, we would be able to recommend the client consider local expansion to Wellington. Then look across to Australia (Perth) for Trans-Tasman Trade. Finally, client could consider the leap into the American market – with various cities suggested in the above table.

6. Further Consideration

The data used for this observation includes high-level quality of life indices which have been used to narrow down cities world-wide that may have similar resident behaviours. Additionally, FourSquare data has been used to identify venues surrounding the city centre to further refine potential focus cities for expansion opportunities.

Ultimately the client should consider their preferred focus countries, then a targeted city investigation could be undertaken. This would delve deeper into city districts and could include further considerations.