Coursera final project

**Report on key findings from clustering analysis of Qld Mining towns**

William Lovell

October 2019

**Introduction**

The mining industry in Queensland, Australia involves many far flung small towns in remote locations, far removed from major cities. This creates many challenges for mining companies to attract and retain employees. If companies had a better picture of the types of businesses/venues available in a small town they could tailor the employee value propositions at different profiles of candidates - for example, a town with amenities suitable for children should be targeted to candidates with families, vs towns with a larger number of facilities for single, younger employees should be targeted at that population.

This project will aim to cluster regional Qld towns by the types of small businesses/venues available in the town to enable mining companies to refine the Employee value proposition and realistic lifestyle previews provided to potential candidates to enhance attraction and improve retention in the long term.

**Data**

Foursquare venue data will be sourced through the Foursquare API.

To identify the mining towns a mixture of sources have been used, including: industry knowledge, the Queensland Resources Council, and this article: <https://www.propertyzest.com.au/2012/10/04/queensland-mining-towns-investment-locations/>

Towns include coal, bauxite, gold, and oil and gas resources and a mix of location including 'outback' towns in the Galilee basin such as Moranbah or coastal towns like Gladstone and Weipa. Latitude and Longitude data has been sourced through industry links.

**Methodology**

Load the packages: I loaded a lot of libraries and packages – including: Numpy, Pandas, Json, geopy, matplotlib, Kmeans from SKlearn, Folium, Geocoder, Lxml and BeautifulSoup4 (not required as data in CSV).

Load the data: the data was sourced from a CSV which included Town, Latitude, Longitude, and Population for 27 towns.

Check and Clean the data: After loading the data is was clear that an additional column named ‘Unnamed:4’ had been loaded so I removed that. I then checked the shape, head, dtypes, and ran a count of towns and that data all looked right. Although I did have to convert Latitude into a Float.

Build a map of QLD: Next I created a Folium Map of the state of Qld so I could review that the data was right and understand the distribution of the towns – e.g. Coastal, Close to city, remote etc.

Get Venue data: after loading Foursquare credentials I then did a search of the top 100 venues within 5K of the first towns data – I then extracted the category of the venues and loaded this into a dataframe. Next we went through all 27 towns and repeated this activity – resulting in the dataframe: Qld\_Venues.

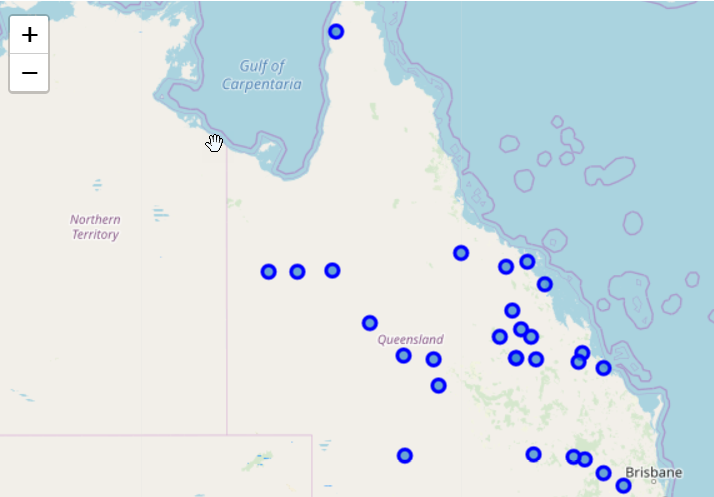
Onehot encoding: To enable clustering and to see which categories were represented in which towns One Hot encoding is used to create unique values for each category – which is finally reduced to the Top 10 categories for each town – meaning a DF with town and top 10 venues for the town to conduct the clustering analysis on.

Clustering: Kmeans package is used to run Kmeans clustering on the towns – 3,4, and 5 clusters were tested with 4 used as this created the most logical grouping of QLD towns (based on # of towns per cluster and local knowledge).

Map and examine the clusters: A Folium map of the clusters was created and cluster details were printed to enable analysis.

**Results**

Map of QLD with towns mapped:

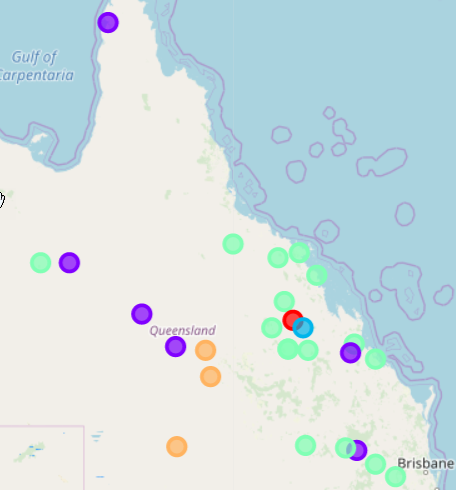


Venues for initial town: the Town was called ‘Toowoomba’ the data returned 84 venues – the top five included a coffee shop, burger joint, breakfast spot, theatre and steakhouse.

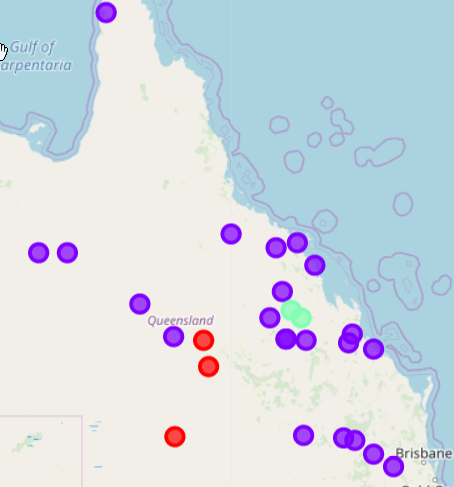
Repeat for each town in data set: This was then looped through for all towns, overall there were 84 categories – Toowoomba had the most venues, with some towns (Emerald, Gladstone, Mackay) having 20-40 venues, and others (Barcaldine, Blackall, Dysart) having less than 4 venues.

Clustering: Kmeans clustering was run on the dataframe and then the clusters mapped to review the results. Clusters were set at 5,3,4 until finally 4 clusters was utilised.

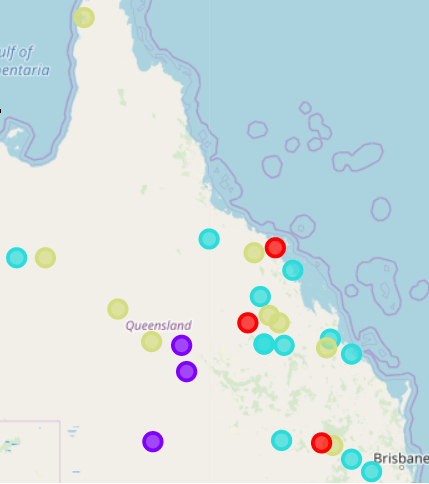
Five Clusters:



Three Clusters:



Four Clusters:



Clusters analysis: The five cluster analysis resulted in a cluster of one town, and the three cluster analysis resulted in one cluster of 21 – whereas the four cluster analysis resulted in a nice mix of towns.

Cluster one: Quilpie, Barcaldine, Blackall . These are simple, small towns highlighted by Airports which indicates they are most likely Fly in Fly out towns, their other venues are simple highlighted by campgrounds, convenience stores, and fast food stores. Populations: 654 – 1316.

Cluster two: Mt Isa, Mackay, Gladstone, Dalby and Toowoomba. The towns in cluster two are larger, and more established with long histories they are highlighted by venues more expected to be found in large towns like Shopping Malls, Supermarkets, and different categories of food options from Fast Food to coffee shops to Italian food. Another feature also appears to the Movie theatres/Multiplexes. This is supported by the population data with towns in this cluster having between 18000 – 96,000 population.

Cluster 3: Weipa, Cloncurry, Collinsville, Chinchilla. Population: 1500-3000. These historic towns appear to have common venues related to tourism including Motels, RV Parks and Campgrounds, Scenic Lookouts, and Concert halls.

Cluster 4: Clermont, Proserpine, Miles Populations: 1169- 3562. The towns in cluster four are geographically aligned and have a mixture of key fixtures including Café’s, Hotels, Grocery Stores.

**Discussion**

The results show interesting clusters showing that there are differences between small towns other then population size. The biggest cluster was related to Population and town size, however the other three clusters where not overly different in town size, or remoteness which indicates that the venue data was finding variance within the clusters – e.g. very simple towns aimed purely at a fly in fly out population, more complex small towns which appear to have a tourism component on top of their mining heritage, and finally a cluster which appear to be a catch all of small towns without those features.

Geographic considerations do not appear to have a large impact with three of the clusters having a mixture of coastal and ‘country’ towns. One feature that might have been nice to include would be a ‘remoteness’ factor. Overall this data does provide an interesting take on small town lifestyles however it is not a complete picture – other features should be included (population, remoteness, some take on economic prospects of the town) however the data can be used to help provide a preview to prospective employees of what living conditions are like the towns.

**Conclusion**

# This project aimed to create a picture of ‘small mining towns’ in QLD to help employers provide their prospective employees with a realistic understanding of the living conditions within the town, including building up a picture of weather a town was ‘family’ friendly or more suited to a single or fly in fly out lifestyle. This was achieved in the analysis however could have included other features to build a more complex picture.