Author: Viet Nguyen

Coursera Capstone Project

Restaurants in Melbourne - The similarity between the inner suburbs.

1. Introduction

a) Description of the problem:

- Running a café or restaurant is not easy. The business owners have to invest a lot of money and time to train their staff as well as buying equipment etc. But no matter how much money invested, choosing a wrong place to start the business ruins everything.

- Hospitality businesses rely on the area where there are demands and competition so much. Where will be the places that suitable to run the business?

- This capstone project collects available data of existing restaurants around Melbourne's inner suburbs then uses a machine learning algorithm to cluster the areas.

- The clusters provide an idea of the potential areas to start running a restaurant in Melbourne in a way that will be neither too specific nor general.

b) Background discussion:

- The chosen city is Melbourne as the author is living here. Melbourne is one of the biggest cities in Australia. And it is famous for the hospitality culture. It is very easy to find a restaurant in Melbourne even you are in its surrounding suburbs. Melbourne people love to dine out. Sounds it is easy to make money by running a restaurant in Melbourne, isn't it? It may not.

- Imagine we run a restaurant with well-trained staff and delicious menu at reasonable prices, but we wanted to save the rent, so we do our business in the area that provides us cheap rental cost (of course, this area is not bustling). We ended up having no customers but still have to pay the bills.

- Or imagine we run a restaurant in a crowded area with so many competitors around us. The rental cost, in this case, is inevitable high. We have to pay for all the expensive costs. Does the chance of getting customers from the same pool as our competitors worth?
Who will be interested?

- As mentioned, definitely the person who is going to run a restaurant business in Melbourne. Choosing the right place to start the business will not only help business owners save a lot of future costs, but they can also gain better revenue and profits.

- Besides that, other parties that rely on the distribution of restaurants such as food delivery riders also may be interested in this project. Knowing where to have more restaurants will bring them more chance of getting trips.

2. Data description:

- The data to be used for displaying the desired results are 2 dataframes. The sources of information from both dataframes are collected through running Foursquare API.
One dataframe holds information of venues found in all suburbs. The other dataframe holds geographical information of each suburb.

- The two mentioned dataframes will then be used for applying K-means clustering algorithm. The algorithm will assign a particular label for each venue.

- With labels defined, it is easy to visualise the suburb's clusters on a map, and also to show data of each cluster such as 'chance of getting customers' and 'competition rate'.

3. Methodology

- This project applied multiple tools and techniques that the author of this report has been trained by Coursera including: IBM Watson cloud service, Foursquare API, Jupyter notebook, Python, Scikit-learn, Numpy, Pandas, K-means clustering algorithm etc.

- Multiple numpy arrays to hold the lists of Melbourne's inner suburbs. This information will be used by Foursquare API to generate a dataframe that holds the suburb's geographical data including postal code, latitude and longitude.

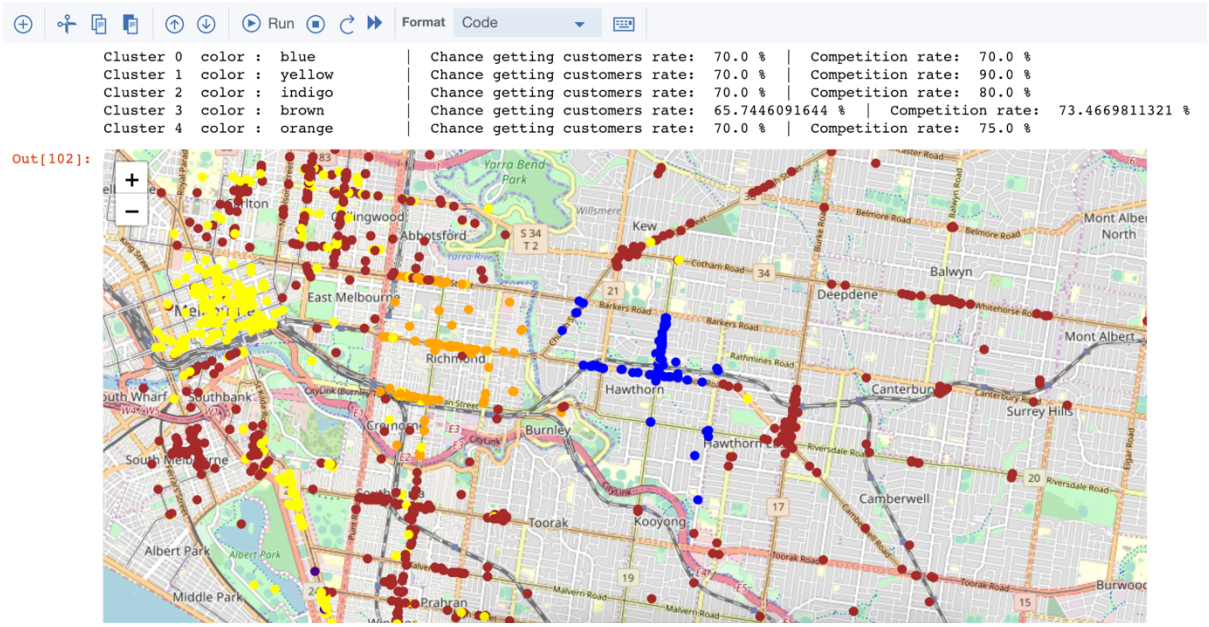
- The generated suburb's geographical data dataframe will then be used by Foursquare API again to generate a dataframe of all venues found that holds information such as venue's suburb, venue's category, venue's location, venue's geographical data etc.

- When the dataframe of all venues is generating, extra information will be added including 'chance of getting customers' and 'competition rate' (author's subjective information based on his real-life knowledge of the areas). This information will be used later by the machine learning algorithm to cluster the areas.

- With generated necessary dataframes, it is possible to apply K-means clustering algorithm with a user's chosen number of clusters (maximum: 12). After running the algorithm, each row in the venues dataframe will be assigned a new label. Each label represents for different clusters.

- All processed data will be together used for generating a map of clustered areas. Different colors for different clusters, information such as 'chance of getting customers' and 'competition rate' of each cluster will also be displayed.

4. Results



- As shown in the image above, there are 5 clusters. Each cluster has a different colour when visualised on the map. Each cluster also has their information about 'chance of getting customers' and 'competition rate'.

- Melbourne's CBD and suburbs close to it are on the left-hand side, and the suburbs that are far from the CBD are on the right-hand side.

- Cluster's details:

+ Cluster 0 (1st cluster): colour: blue | chance rate: 70% | competition rate: 70%

+ Cluster 1 (2nd cluster): colour: yellow | chance rate: 70% | competition rate: 90%

+ Cluster 2 (3rd cluster): colour: indigo | chance rate: 70% | competition rate: 80%
(not displayed on the image above as this cluster stays far below the others. But for your information, this cluster is close to the beach area of Melbourne city called St Kilda.)

+ Cluster 3 (4th cluster): colour: brown | chance rate: 65.74% | competition rate: 74.47%

+ Cluster 4 (5th cluster): colour: orange | chance rate: 70% | competition rate: 75%

5. Discussion

- We can see from the results that the clustering algorithm works really well. Not all areas having restaurants, and not all streets/roads having many restaurants. We clearly able to identify which are places other people running their restaurants. Sometimes, it is good to be

different; but, in this case, it maybe not a wise choice to start running restaurants in areas that nobody wants to do business there.

- The closer to the city centre, the thicker of restaurants distribution. It is definitely correct as people tend to live closer to the city, and in the city, there are a lot of tourists as well. Yellow cluster completely covers the city centre area, that's why it has the highest competition rate with 90% while the chance of getting customer rate is 70%. Definitely not because it has lack of demands, but because the same demands have to be shared for a lot of competitors. Renting costs for areas close to the city centre are expensive as well.

- The brown cluster covers everywhere on the map but not much. It is clearly because it covers mainly areas that far away from the city. This brings down the rate of getting customer just 65.74% as well as the competition rate compared to the yellow cluster at 74.47%. Renting costs of suburbs on the far right-hand side are less than suburbs close to the left-hand side.
- The indigo colour cluster which is not displayed on the image has the second highest competition rate although it covers suburbs far from the city centre, but the areas have a lot of tourists because they are close to the beautiful beaches.

- The blue cluster is interesting as it has the same rate of getting customers at 70% but also has much less competition rate just 70%. It covers the area not too close to the city centre. But the area has a lot of schools, universities.

- The orange cluster is in the between of the city area and the blue cluster's area. Not so many schools or universities or tourists over there. Although it is not in the city centre, but quite close to go there, also parking is way easier. Renting costs in this area are also much less than in the city centre, that's why there are so many restaurants in this area which makes its competition rate to 75%.

6. Conclusion

- The processes to get the results are repetitive processes, the author had to do the same thing over and over again for so many times to get the right outputs. By looking at the results and comparing with the author's real-life knowledge of the areas, the algorithm works perfectly.

- This project gives good analysed outputs to the potential new restaurant owners, helps them having the answer where to run their new business. However, the final decisions are up to them. For example, they know that the blue and yellow clusters have the same rate of chance getting customers but because they want to be in the city centre so they will accept the higher rentals as well as higher competition rate.

- The author prefers the orange cluster most as it is close to the city, rentals are cheaper. Although the blue cluster has the same chance of getting customers rate with lower competition rate, it is too far away from the city centre.