IBM Data Science Professional Certificate (Coursera)

- Applied Data Science Capstone

Project Title: 2 Bedrooms Units' Weekly Rents Index and Surroundings near The University of Melbourne Parkville Campus

Project Report

By: Chai Hoon Lim

April 2018

Contents

I. Introduction

This is the proposal report for the capstone project of IBM Data Science Professional Certificate on the Coursera platform. In this project, I have to come up with an idea to leverage the Foursquare location data to explore or compare neighborhoods or cities of my choice or to come up with a problem that I can use the Foursquare location data to solve.

My idea is to explore areas near to the University of Melbourne Parkville campus, segment and cluster the areas and provide information on the average 2 bedrooms units' weekly rents index for the shortlisted postcode area. This idea arises from the housing need of new students moving to Melbourne to attend courses at the University of Melbourne at the beginning of each semester. For some students choose to rent the whole apartment near to the campus, they need to survey on areas which are near to campus before they can decide which area to check out for rental apartments. With this project, I hope that to assist them in deciding on which areas to look out for rental apartments and that they can find a suitable apartment and settle in before semester commences. And 2 bedrooms units are chosen in the analysis as 2 bedrooms units are the most popular housing options among students.

Even though the idea starts with housing needs for new students moving to Melbourne to attend courses at the University of Melbourne; the targeted place can be updated from University of Melbourne Parkville campus to any places at Melbourne that interested the audience.

II. Data acquisition and cleaning

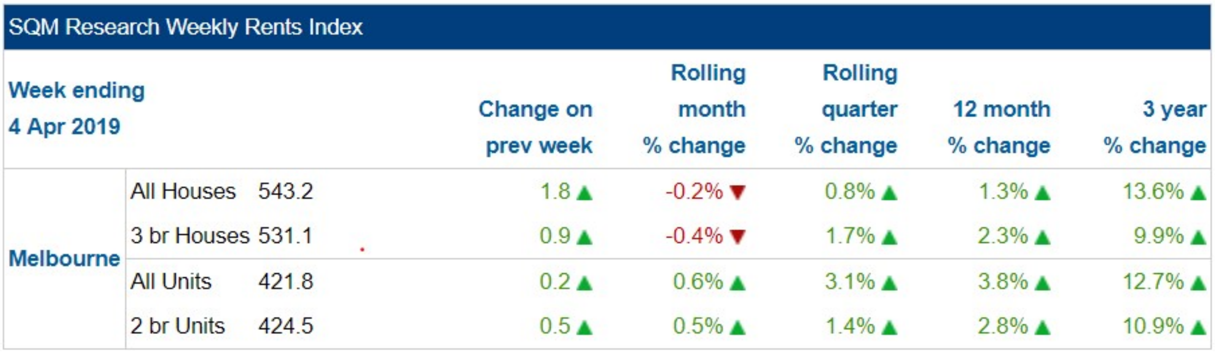
In order to achieve the objective of this project, the following datasets are acquired and cleansed accordingly:

Dataset 1: The coordinate of the University of Melbourne Parkville campus (the targeted place) is used to plot the targeted place on the map.

Dataset 2: List of postcodes for suburbs near to the campus are collected to retrieve weekly rents index as well as for suburbs comparison.

Dataset 3: Average 2 bedrooms units' weekly rents index is scraped from the sqmresearch.com.au using the following query for each postcode in the **Dataset 2**:

https://sqmresearch.com.au/weekly-rents.php?postcode={}&t=1
After that, the result page is parsed to get the "2br Units" rents index from the table similar to the following:



SQM Research Weekly Rents Index

Week ending 4 Apr 2019			Change on prev week	Rolling month % change	Rolling quarter % change	12 month % change	3 year % change
Melbourne	All Houses	543.2	1.8 ▲	-0.2% ▼	0.8% ▲	1.3% ▲	13.6% ▲
	3 br Houses	531.1	0.9 ▲	-0.4% ▼	1.7% ▲	2.3% ▲	9.9% ▲
	All Units	421.8	0.2 ▲	0.6% ▲	3.1% ▲	3.8% ▲	12.7% ▲
	2 br Units	424.5	0.5 ▲	0.5% ▲	1.4% ▲	2.8% ▲	10.9% ▲

Ideally, weekly rents index by suburb should be acquired, but due to no such data available online, weekly rents index by postcode are acquired instead. At the Methodology section, I will discuss this data limitation.

Dataset 4: List of suburbs with locality and longitude and latitude coordinates from the following website:

https://www.matthewproctor.com/Content/postcodes/australian_postcodes.csv

The locality column in the file is translated to suburb name in the program.

Dataset 5: Geo Dataset for the Melbourne area is obtained from the following website that contains each suburbs' coordinates:

https://raw.githubusercontent.com/codeforamerica/click_that_hood/master/public/data/melbourne.geojson

Dataset 6: Foursquare API is used to retrieve venues within 1km radius distance using suburb coordinates.

1. Data Cleansing & Merging

Checking the datasets discover that **Dataset 4** has unmatched suburbs name compare to **Dataset 5** and the suburb name is corrected according to Dataset 5.

Dataset 2, **3** and **4** are merged to get a list of postcode, suburb name, latitude, longitude coordinates, and 2 Br units weekly rents index (by postcode) to create a "Melbourne 2Br Units' Weekly Rents Index" dataframe for each suburb. Each suburb's coordinates are updated using geopy library.

2. How the data is used

Dataset 1 and the merged dataset "Melbourne 2Br Units' Weekly Rents Index" are used to visualize suburbs with 2Br Units Weekly Rent Index superimpose on top the Melbourne map that plot using Dataset 5.

The neighboring venues data in Dataset 6 are used to cluster suburbs into 3 clusters using the k-means clustering algorithm.

III. Methodology

In this project, we focus on studying suburbs near to the University of Melbourne Parkville campus, in terms of the neighborhoods venues and the area's Average 2 Bedrooms Units' Weekly Rents Index (by postcode).

Firstly, we need to get the coordinates of the campus which is available online. Other than that, we need to update the list of postcodes that we are interested to include in this project. The original idea is to use suburbs which are more commonly use when we look for a rental apartment in Melbourne instead of postcodes, but due to no free available weekly rent information by suburb to scrape using a program, postcodes are used instead of suburbs. The rental index is then mapped to each suburb using the suburb's postcode, so suburb with the same postcode will have the same rental index.

Data exploration using a map on the suburb coordinate in the Dataset 4 discovered that some of the coordinates are plot outside the suburb area in the map. To correct this issue, each suburb's coordinates are updated using Geopy library by passing in address in the format "Postcode Suburb State" like "3000 Melbourne VIC". After the suburb coordinates are plot using map and those suburbs with coordinates not in Melbourne area (based on the area coverage in the melboure.geojson) are removed. Data exploration also discover the suburb name for Melbourne is not matching to Melbourne.geojson, the records are updated to Melbourne (3000) and Melbourne (3004) respectively according to the postcode.

Finally, Foursquare API is used to explore each suburb's neighborhood venues within 1km radius distance of the suburb coordinates. Then suburbs are analyzed using the venue categories in each suburb. The occurrence frequency of each venue category is used to group suburbs using k-mean clustering algorithm into 3 clusters. K-mean clustering algorithm is suitable in this analysis as we are interested in finding out the similarity between suburbs and dissimilarity between clusters. Map visualization on the clusters and rental index by postcode, the targeted audience can have some ideas on which suburbs look out for rental apartments.

IV. Discussion

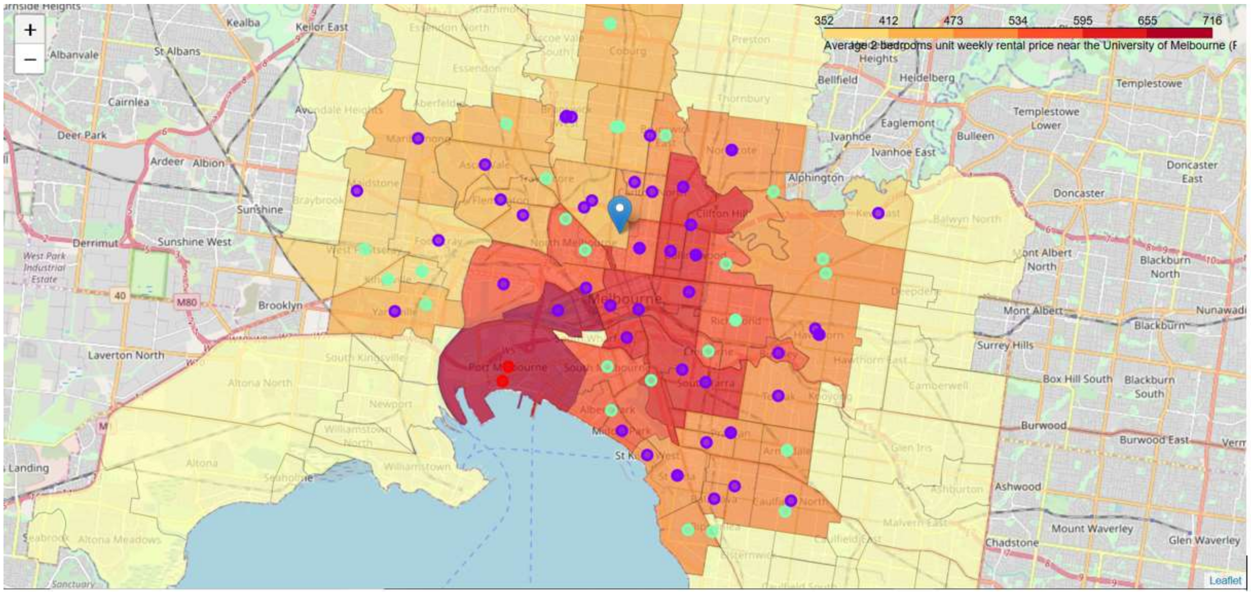


Figure 1: Melbourne map with weekly rents index and suburbs' cluster overlap on top.

The analysis can be more meaningful if the 2 bedrooms units weekly rents index for suburbs near the campus is available for free to scrape using program. To fill the gap of this missing information, the 2 bedrooms units weekly rents index by postcode area is used instead. Postcode area are bigger as each postcode may have more than 1 suburb, so on the map we can observe that suburbs near to each other (they are most likely having the same postcode) are having the same 2 bedrooms units weekly rents index.

While for the suburb clustering, since some of the suburbs are less than 1km radius distance to the nearby suburbs, these suburbs are more likely to group into the same cluster.

Examination on the 3 clusters, we try to determine the discriminating venue categories that distinguish each cluster. Cluster 0 has suburbs with leisure activities venues such as Beach, Go Kart Track, Zoo Exhibit, Soccer Field and Paintball Field. Both Cluster 1 and 2 have cafes and restaurants, and based other characteristics they are areas to live in. Comparing both clusters further, we observe that Cluster 2 has more Pub, Bar, Fastfood Restaurant, Sandwich Place and Pizza Place, and Cluster 1 has more Light Rail Station and Convenience Store nearby.

V. Conclusion

In this project, I analyzed the suburbs near the University of Melbourne Parkville campus to provide basic information to new students which suburbs to look for a rental apartment based on their budget for the rental price, distance to the campus and the suburbs neighborhoods venues. But due to the lack of rental information by suburbs, the information given in the project might not be precise. So to improve this project, rental information by suburbs data should be acquired and use in this project.

Lesson learned in this real project is data might not be available free to acquire, and data always need to cleanse before can be used in analysis. In order to detect data problem, data exploration is important. Also, I applied my knowledge learned in this specialization course and also familiarized with some of the very useful Python libraries.