Where to move

IBM Data Science Professional Certificate

Ricardo Garcia-Rosas

Introduction

One of the most difficult things about moving to a new city on the other side of the world is deciding where to live. While you can look-up online the characteristics of the most famous suburbs many others that could be hidden gems are overlooked. So, what if you could use the place you are currently living in as a reference to look for recommendations of where to live on the other side of the world. This recommendation could include suburb characteristics such as the types of businesses and facilities around it and some statistical information on the rent prices. Thus, narrowing down your search for the perfect place to live on the other side of the world!

This service is aimed at the people of our globalised world, who are looking to move to a new city within their country or internationally. For this project I'll use myself as an example, back in in the day I move from Guadalajara, MX to Melbourne, AU. So, I'll be using the place I used to live in Mexico to rank suburbs in Melbourne that could be suited to me.

Source Suburb:

Providencia, Guadalajara, Mexico.

Target City:

Melbourne, Australia.

Data

Venue Data

Venue data will be obtained using the Foursquare API. For this, it is required to provide the geographical coordinates of the suburbs of interest.

Source Suburb

The location for the source suburb will need to be determined for the Foursquare queries, this can be easily obtained through the Nominatim Geocoder.

This location data will be used to obtain venue information using the Foursquare API. The focus will be in obtaining the top 10 venue categories by frequency in that area.

Target City

Location data for all the Melbourne suburbs is needed for the Foursquare venue queries. This can be easily obtained from the Australian government's data access website: https://data.gov.au/dataset/ds-dga-af33dd8c-0534-4e18-9245-fc64440f742e/details

Similarly to the venue information discussed before, the top 10 venue categories will be required for all Melbourne suburbs. The venue data will be used in the comparison between the two cities (by suburb).

Rent Price Data

Source Suburb

This is probably the most difficult to obtain as Mexico is not great with data gathering. The objective with this data is to obtain the percentile the specific suburb is in w.r.t. rent prices. This is in order to have a better price comparison between the two cities as they may be quite different in terms of cost of living. This is with the assumption that the person is (at least) looking at maintaining their current living conditions.

Thus, the median rental price for housing properties in all the suburbs of Guadalajara, MX are needed.

After a good time searching, I found a database containing rental information for the whole city of Guadalajara at the following link: https://iieg.gob.mx/ns/?page_id=11967

Given that the data in the link contains information for suburbs outside Guadalajara City, only the relevant ones were imported. The dataset contains all available housing properties for rent in the city, thus it is necessary to group and take the means per suburb. After this, it is required to calculate the rent percentile with respect to all the suburbs in the group, this allows for a better comparison between two cities with potentially different cost of living and currency standards.

Target City

Median rental price data for all the Melbourne suburbs can be obtained from the Victorian government's website: https://www.dhhs.vic.gov.au/publications/rental-report

This data will be used to determine each suburb's rent percentile, which will then be used as an additional feature for suburb comparison. The data is provided in weekly median rent prices, so these need to be converted to monthly. Only those for Melbourne will be used. As some suburbs are grouped together in the rental report, these needed to be split.

Methodology

Suburb Features

The venue data was converted to frequency of venue occurrence and normalised. Such that a single number between 0 and 1 represented the frequency (or popularity) of a venue category in each suburb. As not all suburbs may have the same types of venues in their vicinity, the intersection of all these venue categories needs to be used for the comparison. Lastly, the median rent price percentile for each suburb was added as a feature to the data of each suburb.

Suburb Recommendation

The objective of this project is to obtain the top 5 suburbs closest in the target city to the source suburb. To achieve this, the suburbs in the target city were clustered together to find suburbs of similar characteristics. This was done using the K-Means clustering technique. After training the clustering model with the target city data, it was used to estimate what cluster does the source suburb belong to. The top 5 suburbs in the estimated cluster were used for the recommendation.

Results

The cluster estimation recommended a cluster with 4 suburbs, which are in the 37-40 percentile range in rent price. The most common venue in this cluster is Café. They also have a variety of restaurants in their top 10 venues. The recommended suburbs were Armadale, Brunswick, Caulfield, and Chelthenham.

Discussion

The most important item for discussion is how the recommendation is quite accurate! The suburb I currently live in Melbourne is a train stop away from one of the recommended suburbs, and I actually frequent that suburb all the time because of the venues there.

One possibility for improvement, is determining the venue category accuracy of the obtained data. The source suburb venue category results determine that restaurant is the most common type of venue; however, the cuisine is not specified. On the other hand, the target city venue data specifies the cuisine. This could be a shortcoming of the current approach. A way to improve this would be to either combine all the cuisines together in the target city data (which would mean information loss), or explore other sources of information that provide richer data for the source suburb case.

Conclusion

A one to many comparison of suburbs using venue and rental price information was performed to provide a recommendation of what suburb would be suitable for a person to move to. Data was gathered from multiple sources and combined to generate a dataset for the source suburb (person's place of origin) and the target city (the city the person is moving to). A K-Means clustering model was trained using the target city data and then used to estimate what target city cluster the source suburb belongs to. This information was used to provide the recommendation, which, by the author's own experience, was deemed as successful.