IBM Data Science Capstone Project – Melbourne Housing Analysis

Author: Connor G

Contents

[Introduction 2](#_Toc71202548)

[Problem statement 2](#_Toc71202549)

[Data gathering 2](#_Toc71202550)

[Methodology 2](#_Toc71202551)

[Results 2](#_Toc71202552)

[3](#_Toc71202553)

[Discussion 5](#_Toc71202554)

[Conclusion 5](#_Toc71202555)

# 

# Introduction

## Problem statement

For the Data science capstone project, I have chosen Melbourne, Australia as my city of analysis. My stakeholder is a small family that are moving to the city and have a place of work in North Melbourne. With the place of work a constraint the family wants a report on the surrounding suburbs that will include a categorization of popular venues, medium price of housing and the distance to travel to work. Their venues of interest would be cafes and nice parks.

## Data gathering

In order to assist the family as a data scientist I will need to access various datasets and collate them into a simple report to handover to the family to help them decide the optimum location of residence. After some research I will be able to grab relevant data from the below sources:

Suburb location data:

<https://www.matthewproctor.com/full_australian_postcodes_vic>

Medium price data:

<https://discover.data.vic.gov.au/dataset/victorian-property-sales-report-median-house-by-suburb>

I will be able to gather venue data using the Foursquare API and will conduct clustering analysis to group the suburbs by their most popular venues.

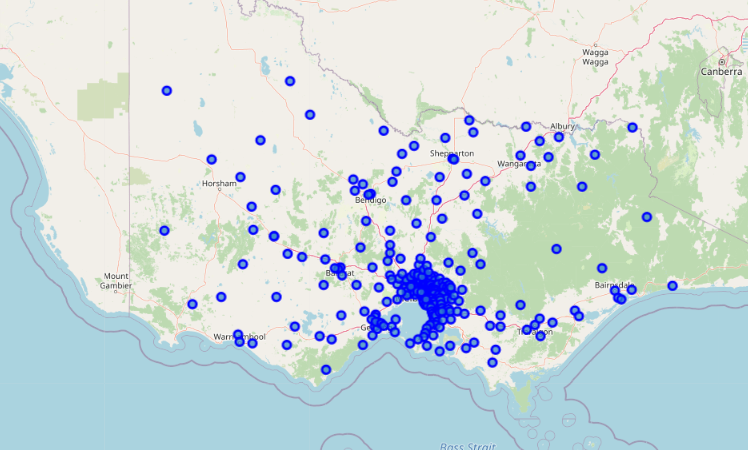
# Methodology

In order to ultimately give the client a suggestion of neighborhoods to investigate, data will need to be gathered and analyzed to ultimately come out with the cheapest, closest to work and most fitting neighborhood for the client. Below are the step taken to achieve the desired solution.

1. Using Python, the above datasets will be imported and filtered to allow easier use later
2. The location data will be used to scrape the venue data using the Foursquare API. This data collection is useful to understand what locations fit the client the best
3. The Venue data will then be clustered using a k-means clustering algorithm to group neighborhoods accordingly and make it easy to select a subset
4. Once a subset is selected that suits the client, the house pricing and distance from work will be analyzed to provide the best fit.
5. Ultimately out of the large set of 320+ suburbs in Melbourne a small subset of roughly 1 – 5 will be selected to present to the client that should fit their needs.

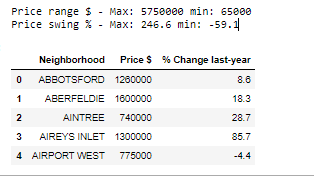
# Results

The results of the experiment proved quite promising and should successfully provide the client with the correct information to make a good decision to purchase their home.

Firstly, looking into the location data of Melbourne, it is quite a large city with many individual suburbs.

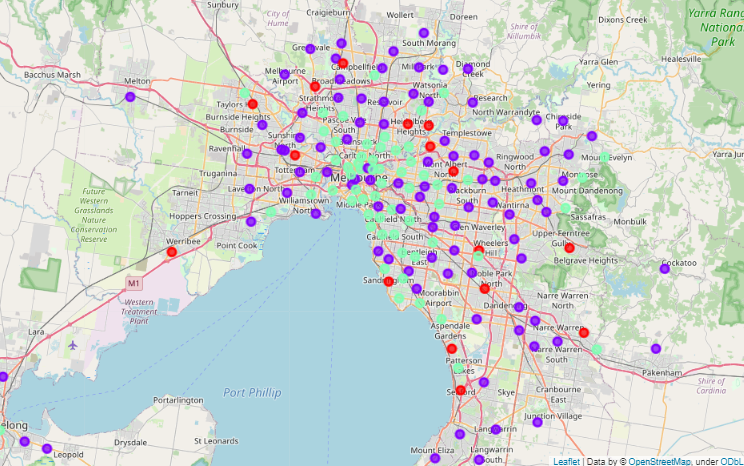
# 

The large size allowed for a very divers data set of venue and price data which will work in our favor when it comes to getting a good fit. The pricing data as below:



Shows a massive range of $5.7m down to $65k and some suburbs experiencing a price increase of more than 240% over the last year.

Finally, the venue data was clustered using K-means clustering. Many different cluster numbers were used to try get the best segregation between neighborhoods and the final number settled was k=3. The reason the cluster size is relatively small is it was able to very clearly define the different categories of neighborhoods. The clusters were then used to colour the location data to visually represent the different venue interests. Given areas near North Melbourne are in interest, a zoomed view is represented.



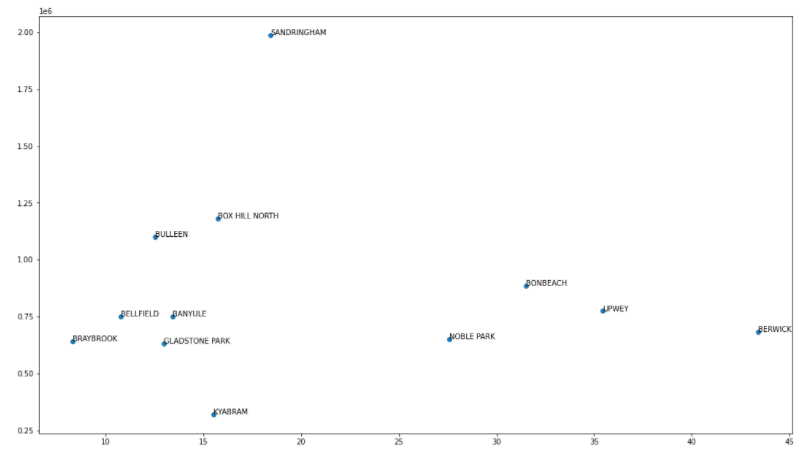
Using the mode of each venue within each unique cluster the below was able to be defined:

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster colour | Venue 1. | Venue 2. | Venue 3. |
| Purple | Japanese | Cafes | Burgers |
| Red | Parks | Bakeries | Grocery stores |
| Green | Cafes | Hotels | Pubs |

Given the clients constraints being Parks and restaurants our clear cluster of interest would be the red cluster with the possibility to try and have it close to some purple clusters. Without making further assessment, above are 2 circled clusters of the red group that could serve well for the client given a proximity to work and well centralized with their interests.

Finally, now that there is a subset of suburbs that suit the client (red cluster) the distance and median price data was added and a scatter plot created (see below). Given their small distance to work and low price, the below suburbs would be a great fit for the client

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Neighborhood** | **Price $** | **% Change last-year** | **Distance KM** |
| **0** | BRAYBROOK | 642000 | -1.1 | 8.3213 |
| **1** | BELLFIELD | 750000 | -11.2 | 10.7731 |
| **2** | BULLEEN | 1100000 | -8.3 | 12.5217 |
| **3** | GLADSTONE PARK | 632000 | -1.6 | 12.9701 |
| **4** | BANYULE | 750000 | -11.2 | 13.4360 |
| **5** | KYABRAM | 320000 | 22.4 | 15.5253 |
| **6** | BOX HILL NORTH | 1180000 | -3.5 | 15.7333 |
| **7** | SANDRINGHAM | 1987500 | 21.5 | 18.4301 |
| **8** | NOBLE PARK | 650000 | 2.7 | 27.5764 |
| **9** | BONBEACH | 885000 | -1.7 | 31.5112 |
| **10** | UPWEY | 774500 | 4.4 | 35.4503 |
| **11** | BERWICK | 682500 | -1.4 | 43.3913 |



# Discussion

The outcome of the data analysis is positive that using venue and pricing data neighborhoods of different countries should be quite easily categorized and sorted to fit specific demographics and price points. For the Melbourne example in this report, out of a total of 328 suburbs a subset of 4 suburbs was extracted that will fit the needs of the specified client. In order to further disaggregate this dataset, the distance between significant venues (parks, restaurants) could also be added into the dataset, i.e biasing suburbs that not only contain such venues but are also close to the venues. With enough data it could be possible to almost select to best fit suburb without even traveling there and instead just using online resources.

# Conclusion

The investigation started with a problem statement that provided the constraints that a suburb must be located with the condition that it is near park and restaurants. The suburb would also be a better fit if it were reasonably priced and placed close to work in North Melbourne. With this problem statement an investigation took place, and all 328 suburbs of Melbourne were taken for consideration and finally 4 were selected for the client. In order to further disaggregate this dataset, the distance between significant venues (parks, restaurants) could also be added into the dataset, i.e biasing suburbs that not only contain such venues but are also close to the venues. With enough data it could be possible to almost select to best fit suburb without even traveling there and instead just using online resources.