The Battle of the Neighbourhoods

Contents

[1.0 INTRODUCTION 2](#_Toc14962983)

[1.1 Background 2](#_Toc14962984)

[1.2 Problem 2](#_Toc14962985)

[1.3 Audience 2](#_Toc14962986)

[2.0 DATA 2](#_Toc14962987)

[2.1 Data Sources 2](#_Toc14962988)

[2.2 Data Cleaning & Feature Selection 3](#_Toc14962989)

[3.0 METHEDOLOGY / DATA ANALYSIS 3](#_Toc14962990)

[3.1 Housing Price 3](#_Toc14962991)

[3.2 Explore Neighbourhoods 5](#_Toc14962992)

[3.3 Questions 5](#_Toc14962993)

[4.0 RESULTS 5](#_Toc14962994)

[5.0 DISCUSSION / RECOMMENDATION 7](#_Toc14962995)

[6.0 CONCLUSION 7](#_Toc14962996)

# INTRODUCTION

## 1.1 Background

Melbourne, the capital of the Australian state of Victoria and the second-most populous city in Australia and Oceania, is a popular immigration city. Many people come there as new citizens every year. With the help from Data Science, they can make better housing decisions.

## 1.2 Problem

When people move to a new city, there are a lot a of obstacles and difficulties in their ways. One of the main challenges is to find a suitable neighbourhood. Each neighbourhood is suitable for some groups of people according to their preferences and financial ability. In this project, I would suggest a neighbourhood to a user by asking them five multiple choice questions and how much they can afford to pay for housing.

## 1.3 Audience

Therefore, this project aiming at providing guidance for people who are looking for a house in Melbourne, to find the right neighbourhood to live in according to his/her financial situation and preference of neighbourhood atmosphere.

# DATA

## 2.1 Data Sources

Two main sources of data are applied for this project:

• Foursquare location data: Foursquare is a local search-and-discovery service. it features a developer API that lets third-party applications make use of Foursquare’s location data. I use this API to search and find out about different venues and their categories of each neighbourhood.

• [Melbourne Housing Snapshot](https://www.kaggle.com/dansbecker/melbourne-housing-snapshot): I download this dataset from Kaggle.com to understand the price of houses in Melbourne. This dataset contains a lot of features but I only use the ‘price’ to normalize housing price and location of the house.

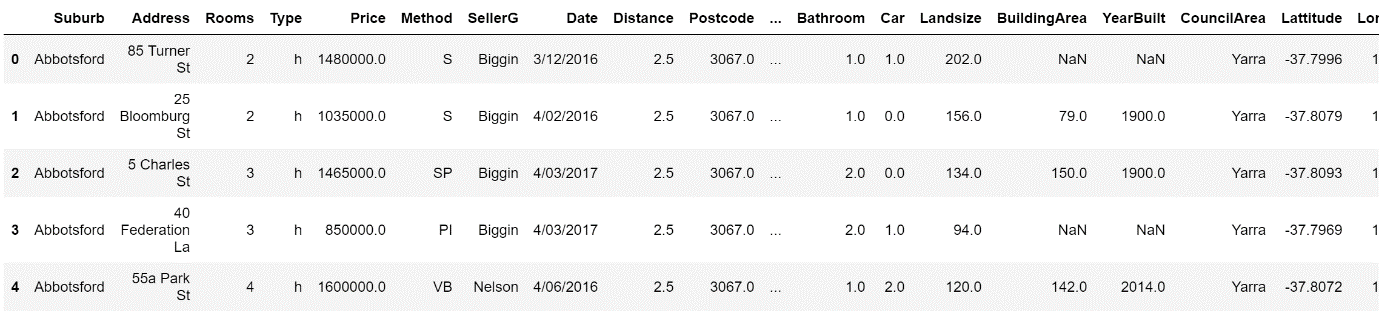


Table 1: Melbourne Housing Dataset

## 2.2 Data Cleaning & Feature Selection

To simplify this problem, we do not consider rural areas, only metropolitan areas are investigated. In order to find out what is the location of each neighbourhood, the dataset is grouped by region and council area.

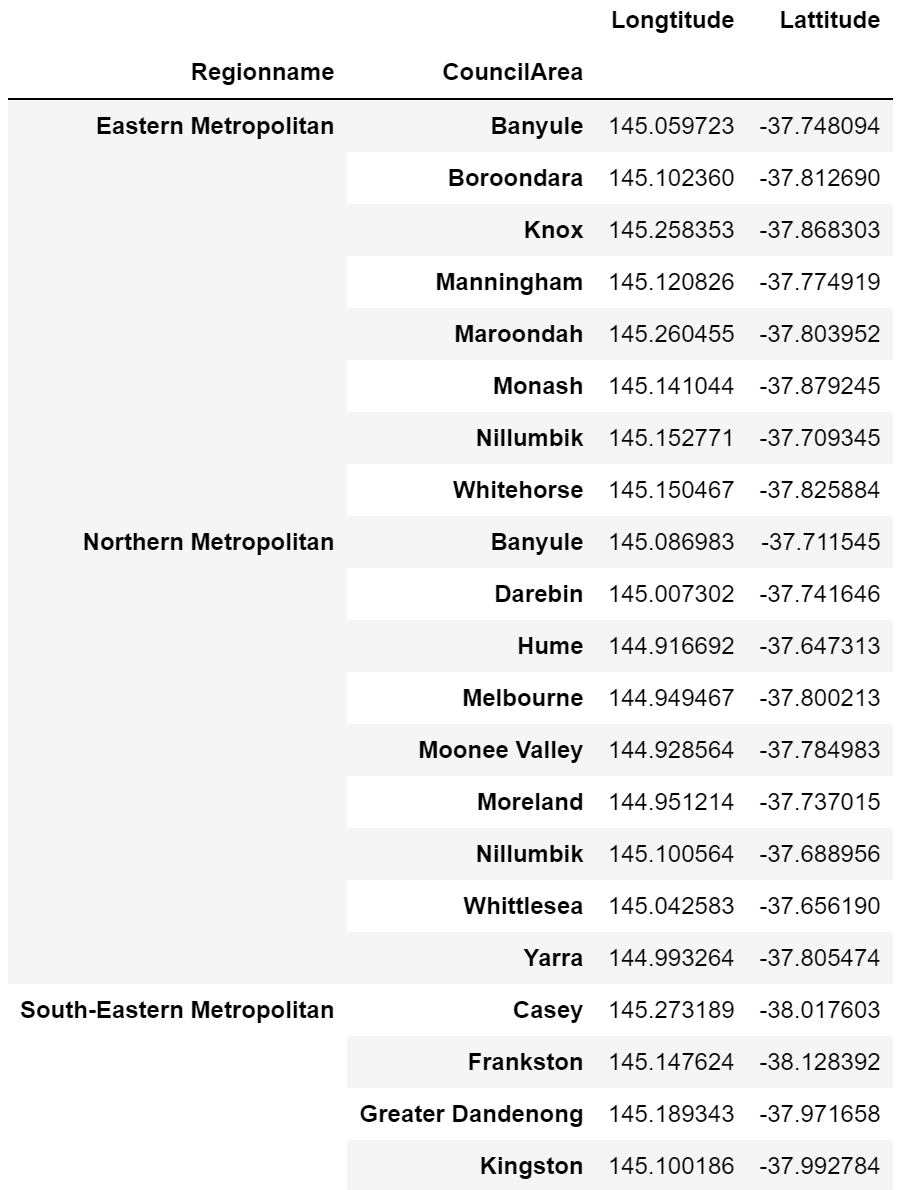


Table 2: Grouped Dataset

# METHEDOLOGY / DATA ANALYSIS

## 3.1 Housing Price

First of all, each region name in above data frame is visualised with different colours. The mean price of each neighbourhood is calculated to compare the housing price, the result is shown in Table 4.

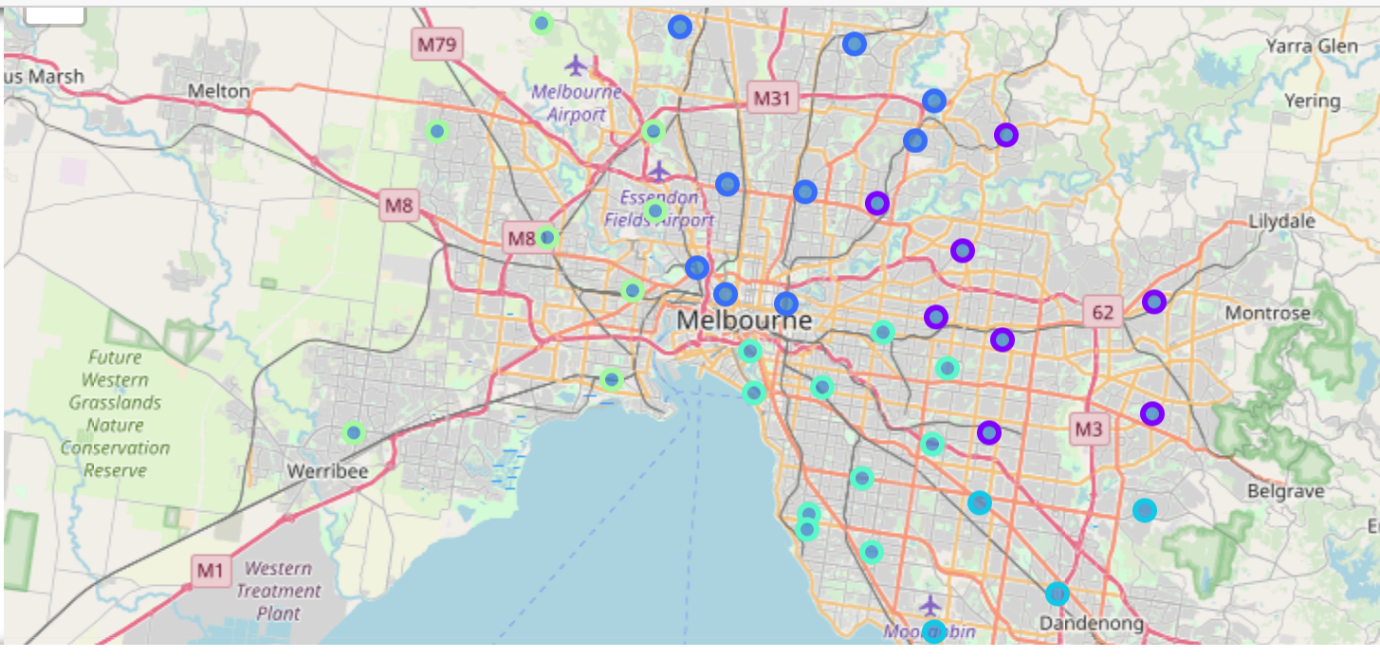


Figure 3: Different Neighbourhoods



Table 4: Mean Price

## 3.2 Explore Neighbourhoods

Now we will build a data frame that contains information about venues of each neighbourhood in Melbourne using Foursquare API. The table below shows the venues in Melbourne:

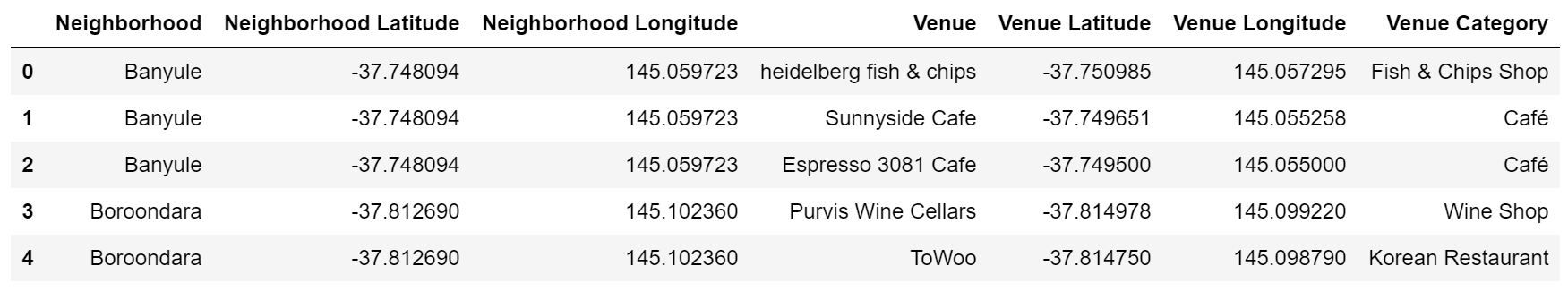


Table 5: Venues

## 3.3 Questions

The content-based recommendation system is used in this project. Users answer to five question about their preferences, and their profile will form according to their answers. Each question asks user “Which kind of place do you prefer to go?” and user can choose between four different options. For example, choices are, ‘Chinese Restaurant’, ‘Indian Restaurant’, ‘Falafel Restaurant’, and ’Korean Restaurant’. Also, I ask user how much money they can afford for housing.

I had no other option than content-based recommendation system, because the system does not have other users’ information, so collaborative filtering recommendation system is not feasible. In the final step, after using user’s profile and the matrix of the neighbourhood to rank every neighbourhood according to user’s profile, I choose a neighbourhood with highest rank between neighbourhood that user can afford.

One important note is if we want to know our users as best as possible with five questions, it’s important to ask user good questions. I try to reach that goal by using two well-known feature selection techniques: Low Variance Filter and High Correlation Filter.

Low variance filter logic is features with low variance are less informative. I choose 60 categories with highest variance at first step. High correlation between two variables means they have similar trends and are likely to carry similar information. So, for each question if choices have low correlation with each other our question is more meaningful. I choose threshold of low correlation equal to 0.6.

# RESULTS



Table 6: Questions and Choices

I demonstrate my result with an example. Table 6 above represent an example of five question that we choose to ask user. Table 7 below shows a user profile according to her answers to above questions, the last row represents how much the user could afford for a house.



Table 7: User Profile

Finally, the best neighbourhood for this user to live can be generated, which is *Melton*. Figure 8 shows perfect neighbourhood for this user with red circle on the map of Melbourne.

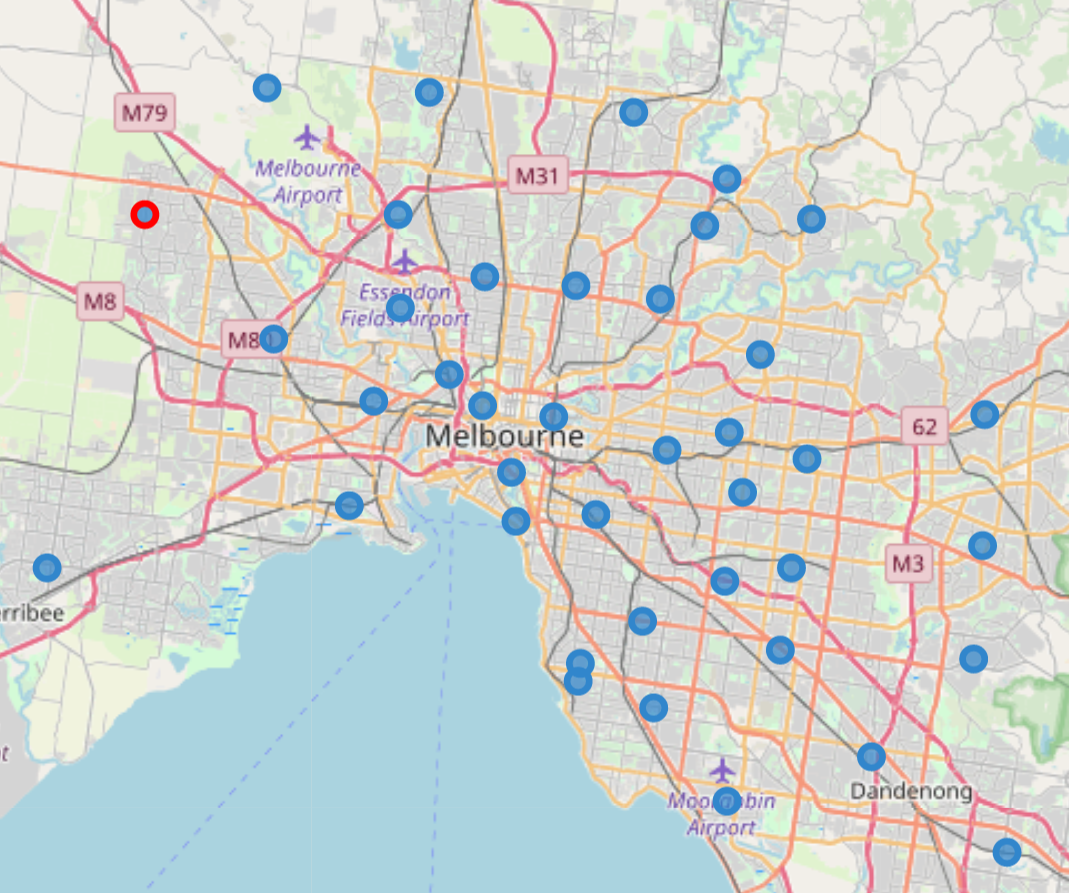


Figure 8: The Chosen Neighbourhood (Red Circle)

# DISCUSSION / RECOMMENDATION

* This project can be further improved by engaging more features like rent price, distance from point that user can choose, rate of crime, etc.
* We can also take rural areas into consideration, to make the results more well-rounded, and fit different users’ requirements.
* This project can be adapted to other cities as well. For example, use Paris’s data instead for people who plan to move to Paris.

# CONCLUSION

In this project, location data of venues and average housing prices of neighbourhoods in Melbourne are used to find the perfect neighbourhood for any user. This content-based recommendation system would greatly facilitate people’s decision-making in terms of housing.