House Hunting in Perth

Using k-Means Clustering to Find Your Next Home

Introduction

When looking to buy a house, location is one of the most important factors people need to consider. House hunters want to live somewhere that matches their lifestyle and personal needs. They also want somewhere that is safe and importantly, within their price range. Embarking on a search to buy a house can be a frustrating task. Australian real estate websites often let you enter a maximum of 5 suburbs to search at any one time, and focus mainly on filtering for price and physical features such as number of bedrooms and bathrooms etc. As for choosing a location, house hunters and property managers often do this based on familiarity with a certain area or by what kind of things (good or bad) they may have heard about a place. Very rarely do buyers have access to a central repository of neighbourhood characteristics as well as historical pricing data and the time and effort it would take for them to research all of this is even more rarely an option. In a world full of data, why are buyers still left to their own devices when it comes to choosing where to spend their lives?

Data

In order to assist prospective buyers to find the right place to call home, a map will be designed to provide information on each neighbourhood of Perth, including the most popular venues and current housing prices. Foursquare venue data will be utilised to determine the most popular venues of each suburb. A classification algorithm will then be used to group neighbourhoods based on their top venues, thereby creating a neighbourhood profile such as 'Café Haven' or 'Shopping Oasis', which will be displayed as markers on a map. Users can read short descriptions about each profile and selectively filter neighbourhoods on the map based on their lifestyle preferences. Median housing prices of Perth will also be displayed on the map so that users can get a general idea of whether a particular area is in their price range, as well as see how this neighbourhood's median house price compares to the rest of Perth.

The data sources to be used include:

Australian Neighbourhood and Postcode data

Foursquare Venue Data – restricted to neighbourhoods in Perth, Western Australia

WA Market: Perth Suburbs Price Data provided from REIWA

Methodology

All data cleaning, transformation and analysis was performed using a Jupyter Notebook provided by IBM Watson Studio.

1. Data Acquisition/Cleaning/Transformation

Neighbourhood Data

A list of Australian neighbourhoods and postcodes with corresponding latitude and longitude coordinates was obtained from https://www.corra.com.au/australian-postcode-location-data/ (file name Australian Postcodes CSV file – Zip). The neighbourhoods were restricted to the Perth region of Western Australia (postcodes in the range 6000-6199). The geopy library was used to get the exact coordinates of Perth to be used when creating the map.

Venue Data

A list of venues and venue categories for each neighbourhood was obtained using the Foursquare API's explore function. One hot encoding was used to create binary variables for each venue category and then rows were grouped by neighbourhood to get the mean frequency of occurrence of each category.

Median House Prices Data

Median house prices were obtained from https://reiwa.com.au/the-wa-market/perth-suburbs-price-data/. The average of all median house prices in Perth was calculated and then used to calculate the percentage difference of each neighbourhood's median price from this average.

2. Clustering the neighbourhoods by venue

The optimal number of clusters was determined by plotting an elbow curve of number of clusters used and related inertia. The optimal number of clusters appears to be somewhere between 4 and 6, as per below figure. To offer users a more personalised experience, 6 clusters were chosen.

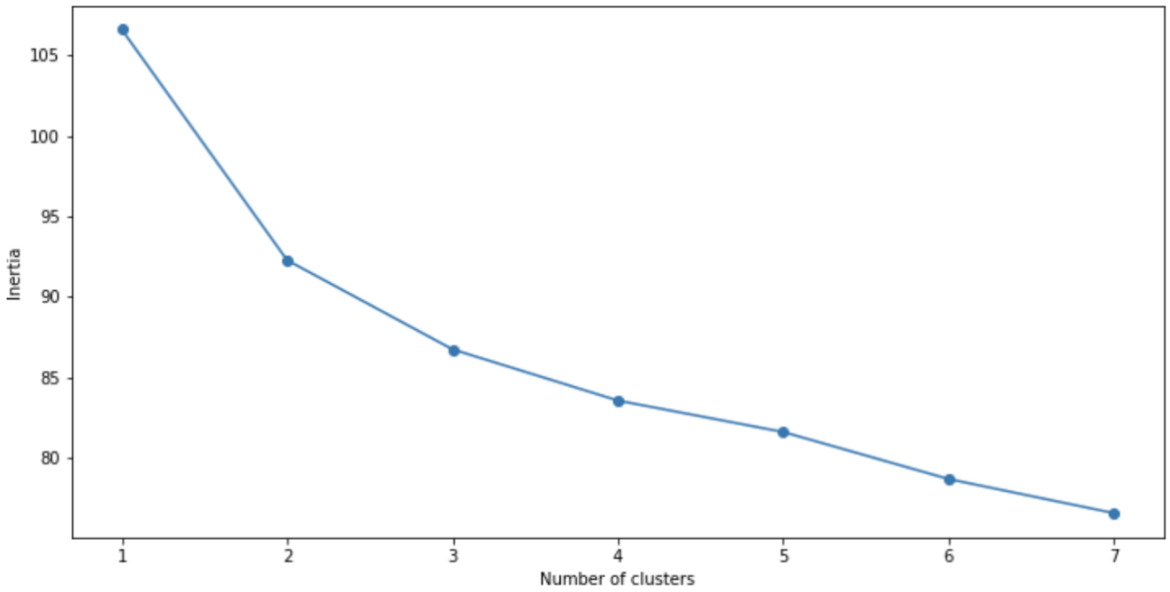

Figure 1. Determining the optimal number of clusters using the Elbow method

k-Means clustering was used to cluster neighbourhoods based on the relative frequency of each venue category. The top 3 most common venues of each cluster were then analysed and descriptive labels were assigned based on these venue categories.

3. Creating the map

The venue data (including top venues and cluster names) was joined to the median house price data. The folium library was used to plot each neighbourhood as a marker on a map of Perth. Each marker was given a label that includes neighbourhood name, cluster name, median house price and percentage difference of median house price from Perth average.

Results

Grouping the neighbourhoods into 6 clusters resulted in the summary shown below, which includes the top 3 most common venues for each cluster.

Cluster	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	Cluster Name
1	Playground	IT Services	Farm	**Family Friendly**
2	Cafe	Yoga Studio	Farm	**Coffee & Yoga Enthusiasts**
3	Park	Yoga Studio	Indie Movie Theatre	**Fitness & Entertainment**
4	Shopping Mall	Grocery Store	Grocery Store	**Shopping Oasis**
5	Cafe	Cafe	Cafe	**Café Haven**
6	Fast Food Restaurant	Fast Food Restaurant	Grocery Store	**Convenient Living**

Presentation of the clustered neighbourhoods on a map showed good distribution of each cluster across the Perth region, as shown below.

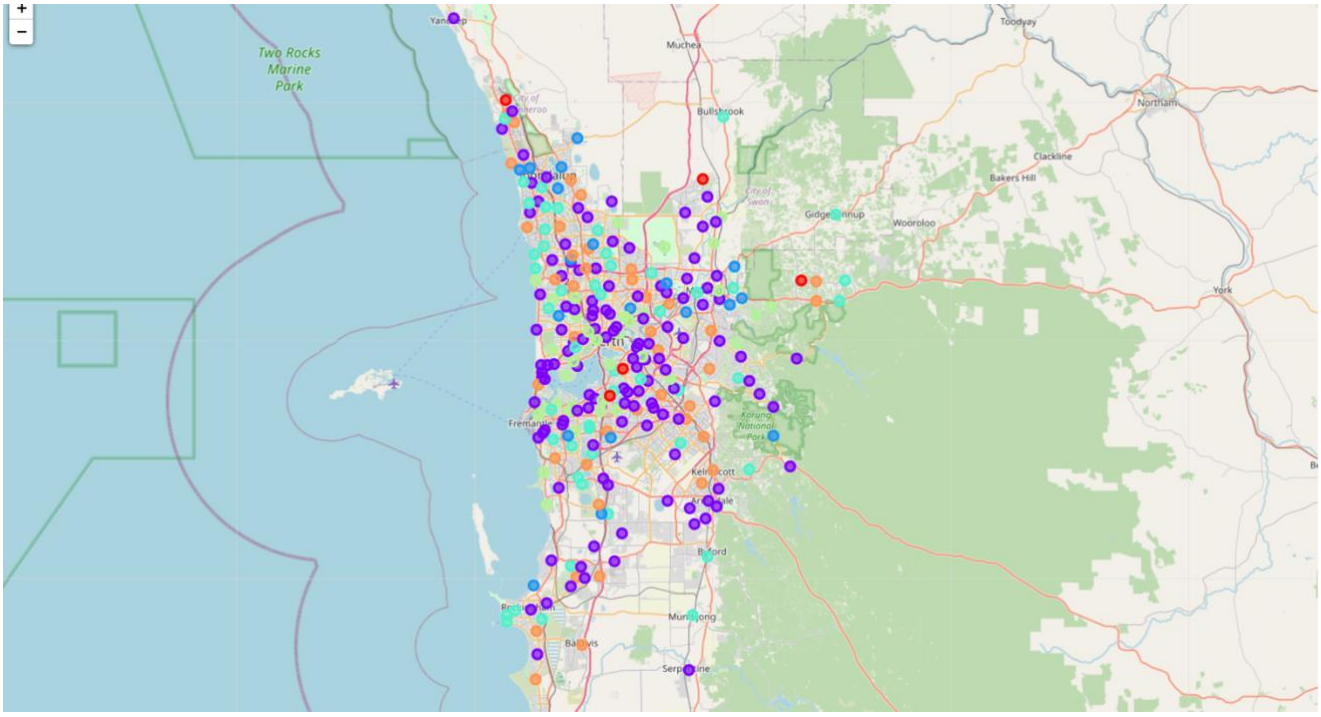

Figure 2. Map of Perth with neighbourhoods clustered

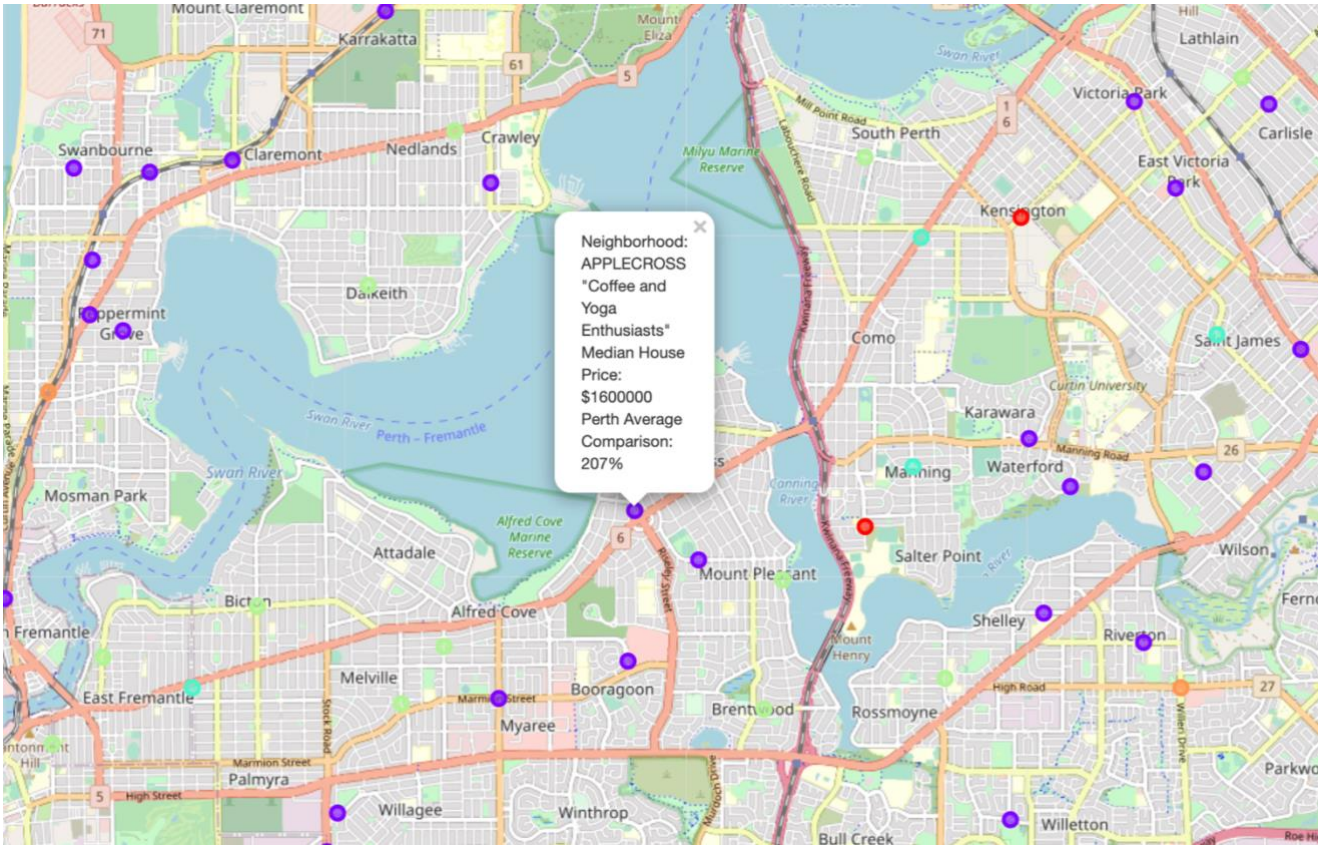

Figure 3. A close up view of the map showing an example label

Discussion

The resulting map provides users with an indication of the type of neighbourhoods in Perth, as well as the median house prices of each neighbourhood. Whilst this may be very useful for some house hunters, they are likely to benefit from additional information such as specific venue names, which type of venues are most highly rated in a particular neighbourhood and relative proximity of these venues to a specific address they may be looking to purchase a house at.

Ideally, this map would be made available to users via website or mobile application and would include filter functionality to allow them to specify the type of neighbourhood, house price ranges or even specific neighbourhoods themselves. This map could also be linked to certain real estate websites, allowing users to 'drill through' to retrieve a list of the current houses for sale in any given neighbourhood.

Conclusion

House hunting is often a stressful experience due to the lack of a central repository of information about neighbourhoods, including neighbourhood characteristics and median house prices. With a map like the one created here, users have a way to easily visualise all neighbourhoods of their city and get quick information that can help them narrow down their house hunting search. Although beneficial as it is, this map has incredible potential to be an essential tool in the real estate industry if it were designed to include more detailed information on neighbourhoods, as well as provide lists of available properties from real estate websites.