IBM Capstone

What is the best place to open a restaurant in Canberra, Australia?

# 1 Introduction

## 1.1 Business problem

The aim of this project is to consider the city of Canberra in Australia and identify suitable locations to establish a restaurant. The suitability of the location will be sourced clustering each neighbourhood and extrapolating the popular places for a restaurant to be established.

## 1.2 Background

Canberra is Australia’s capital city. Despite being the official capital, its relatively small compared to the major cities talked about like Melbourne. Compared to Melbourne, the city of Canberra has a fraction of its population, population density, and geographical area. However, the city is home to several important features and monuments such as the Parliament House, Major Universities, Royal Australian Mint, Australian Institute of Sport, National Gallery, etc. The city has experienced strong and sustained growth, adding approximately 30,000 people since 2012 (act.gov, accessed 2020).

## 1.3 Interest

Canberra is a planned city which should scale well into the foreseeable future as Australia’s population grows. With its future potential, and important significant, Canberra is a strong location to consider when establishing and operating a new restaurant.

# 2 Data

## 2.1 List of Data Sources

1. Mathew Proctor Australian Postcodes: <https://www.matthewproctor.com/australian_postcodes>

2. Foursquare Places API: <https://developer.foursquare.com/docs/places-api>

## 2.2 Data Acquisition

For this project, data is acquired to process geographical location data and leverage the Foursquare API to map all listed venues in the area with their associated category. A feature set containing the name of suburbs of a given postcode and coordinates for each postcode is to be used to leverage Foursquare. A mathewproctor.com has access to csv files which contain postal information by state. The datasets contain the postcode, state, locality, coordinates, and other postal information. The important features from this dataset are postcodes, state, locality, and coordinates. Foursquare will be primarily utilising the coordinate information from the postcode dataset, and the extrapolated venues will be used to cluster each locality.

## 2.3 Postcode Data Overview

**Figure 2.3.1: Mathew Proctor ACT Postcode Data Snippet**

![Table

Description automatically generated]()

The postcode data is in tabular format. The necessary features for this project are Postcode, Locality, Longitude and Latitude.

## 2.4 Foursquare Overview

The foursquare API provides access to a global database of venues. The API was used in this case to search for all venues in the surrounding areas which were gathered from the latitudes and longitude coordinates for each postcode.

## 2.5 Data Selection/Cleaning

The primary data used for this project is referenced in figure 2.3.1. The cleaning process of this data included removing unnecessary features from the table, removing null values from the dataframe, and grouping all suburbs which had the same postcode. The cleaning process is shown in the following figures:

**Figure 2.5.1: Initial dataframe creation**

![Graphical user interface, text, application, email

Description automatically generated]()

An initial dataframe is created from the Mathew Proctor dataset. The extracted dataframe only includes locations within the state of ACT, where Canberra is located as that is the only area of interest for this project.

**Figure 2.5.2: Null values**

![Table

Description automatically generated]()

All null values in the dataframe are filtered. Since there are only 4 datapoints which are null, they can be removed without significantly affecting the sample size.

**Figure 2.5.3: Flatten dataframe**

![Graphical user interface, text, application, email

Description automatically generated]()

The dataframe is flattened, with all locations which share the same postcode being combines into single datapoints. This is done because they have the same longitudes and latitudes and as such, would make mapping all points as individual locations confusing.

**Figure 2.5.4: Get Lat and Long**

![Graphical user interface, text, application, email

Description automatically generated]()

Latitudes and longitudes are extracted from the postcodes and appended into the table as new columns. The latitudes and longitudes are used for this project as values given to the Foursquare API.

**Figure 2.5.5: Final null drop**

![Graphical user interface, text, email

Description automatically generated]()

As a final step to the data cleaning process, null values from the lat and long columns are dropped to prevent errors with the API data retrieval function.

After the total cleaning process, the dataframe consists of 29 rows and 4 features.

# 3 Methodology

## 3.1 general process

The overall process of answering the business problem for this project is to identify how venues within Canberra are clustered together. For the purposes of this exercise, the relative density of similar businesses in each area will be considered for determining where the optimal place is to open a restaurant.

**3.1.1: Folium map**

![Map

Description automatically generated]()

Folium is used as the primary visualisation tool for this project. Folium is typically used to generate maps which zoom in on specified coordinated and markers can be placed on specified coordinates within the map. For this project, an initial folium map is generated to show the Canberra region and all relevant postcode datapoints discussed in section 2.

**3.1.2: Foursquare usage**

![Graphical user interface, table

Description automatically generated]()

The Foursquare API is used to extract all venues and their categories from the coordinate data.

Venues are then sorted by their respective frequency in each area and sorted into a new dataframe.

**3.1.3: Venues sorted**

![Table

Description automatically generated with medium confidence]()

## 3.2 K-means clustering

K-means clustering is an unsupervised machine learning algorithm which is used to group datapoints together into clusters of similar qualities.

K defines the number of groups for all datapoints to be clustered into. The optimal value for K depends on the size of data and its variation. The optimal K can be identified using the elbow method, which involves running the k-means clustering algorithm multiple times with different K values and the results are plotted in relation to their value and score.

**Figure 3.2.1: Elbow method**

![Chart, line chart

Description automatically generated]()

From the elbow method shown in figure 3.2.1, it can be shown that the slope starts to shift at approximate 6 clusters. Therefore, K should be 6 for this project.

Finally, the sklearn clustering algorithm is used with a K value of 6 and all areas are clustered together according to their venue frequency.

**Figure 3.2.2: Areas clustered**

![Table

Description automatically generated]()

# 4 Results

**4.0.1: Folium map of clusters**

![Map

Description automatically generated]()

The folium map above shows each datapoint being distributed into a cluster. Each cluster is coloured differently.

**Figure 4.0.2: Cluster 1 most frequent venues**

![Graphical user interface, text, application

Description automatically generated]()

Figure 4.0.2 shows a breakdown of cluster 1. Each value is weight based on their position from figure 3.2.2.

# Discussion

Areas clustered into cluster 1 were the most notable in relation to opening a restaurant. The makeup of businesses in cluster 1 areas mainly consisted to restaurant/food type venues. Cafes are the most common venues in cluster 1, followed by supermarkets and including stores and various styles of restaurants.

# Conclusion

Given the results of this project, a potentially location for a new restaurant in Canberra will be in one of the regions of Cluster 1. Give its low area and population size, venues are fairly spread around the city. Further analysis should be directed towards the specific population density within each cluster 1 zone. That being the case, Woden may be a reasonable area to open a new restaurant due to the proximity of cluster 1 areas.

# References

A growing and diversifying population - ACT Planning Strategy. (2021). Retrieved 12 May 2021, from https://www.planning.act.gov.au/act-planning-strategy/setting-the-scene/a-growing-and-diversifying-population