Applied Data Science Capstone Project

By Anthony Quast

# Introduction

The Australian start-up company "Bulk like Hulk" is looking to expand their online supplement shop by setting up 10 stores in Melbourne. Due to the nature of the business, it would be desirable for the stores to be located in suburbs which contains as many gyms / fitness studios as possible, as its target audience largely consists of 'gym-goers'. The stakeholders have requested, due to the naming of the shops, that no two shops can be located in the same suburb.

Therefore, the aim of this project is to provide an educated recommendation for the locations of the stores, enabling the stakeholders of "Bulk like Hulk" to choose suitable suburbs which will maximise the profit of their business.

# Data

Two sets of data were used for this project. Firstly, Matthew Proctor's 'Australian Post Codes' dataset (<https://www.matthewproctor.com/australian_postcodes)>, which contains all Australian postcodes along with their locality, state, latitude and longitude (eg: 3144 - Malvern - VIC - 145.034 - -37.857 ...). It was skimmed for postcodes that are within 10km of the city centre of Melbourne, which allowed for the relevant data to be retrieved from the Foursquare database. This subsequently provided gyms/fitness studios for each suburb, which was analysed to draw the right conclusions.

The Foursquare database was skimmed for venues in each suburb by passing a request with the ‘query’ parameter set to ‘gym’. The resulting set of data was refined by further specifying the ‘venue category’ column, in order to remove irrelevant categories such as ‘gym pools’. Duplicates were also removed, as some venues appeared twice. The resulting dataset contained 73 rows, each of which had a neighbourhood, neighbourhood latitude, neighbourhood longitude, venue name, venue latitude, venue longitude, and venue category specified. All of these were used for further analysis.

# Methodology

## K-means analysis

In order to segregate the locations of the gym dataset into 10 separate clusters, the k-means package from the scikit-learn library was utilised. The means of the clusters were found, which provided feasible locations for the new shops to be set up. Using the folium library to generate a map of melbourne, the locations of the gyms (colour-coded to show which cluster each belongs to) were visualised. Subsequently, the cluster means were also visualised to show the 10 recommended locations.

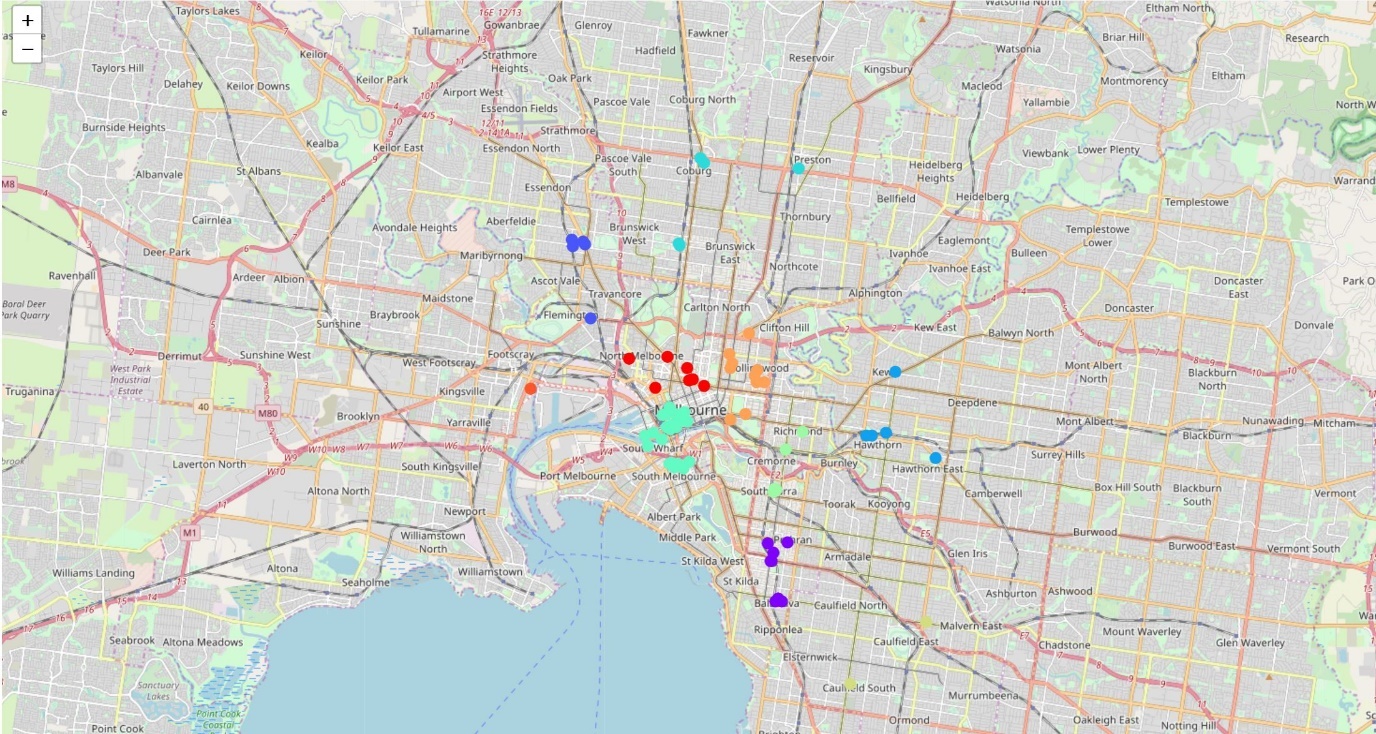
## Finding the modal suburb

Using simple exploratory data analysis, the suburbs which contained the highest number of gyms were found. The dataset was simply grouped by suburb, and then sorted by the # of venues for each. With the use of folium, the resulting suburbs were visualised on the map of Melbourne, displaying feasible locations for ‘Bulk like Hulk’ to set up shops.

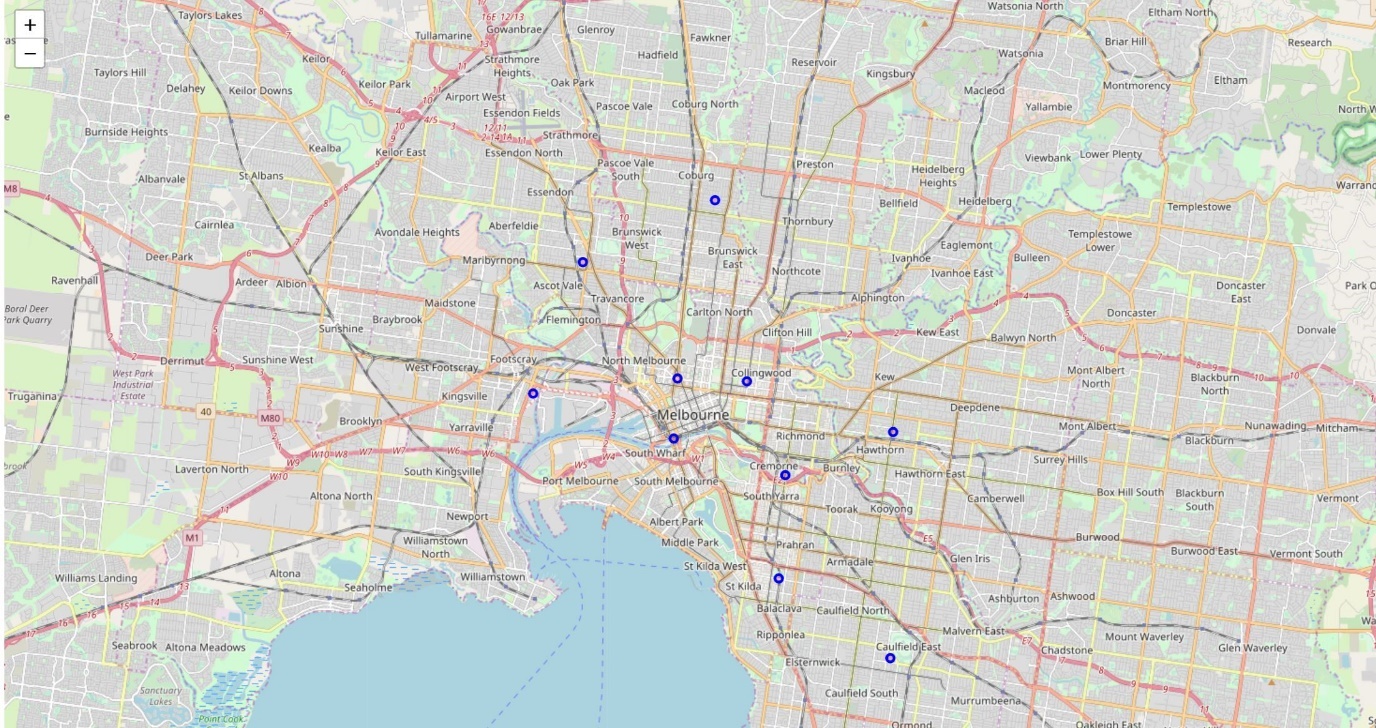
# Results

## K-means

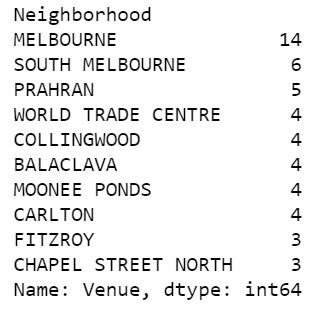
The following map displays the location of each gym on the map of Melbourne, colour-coded to visualise which cluster each venue belongs to:



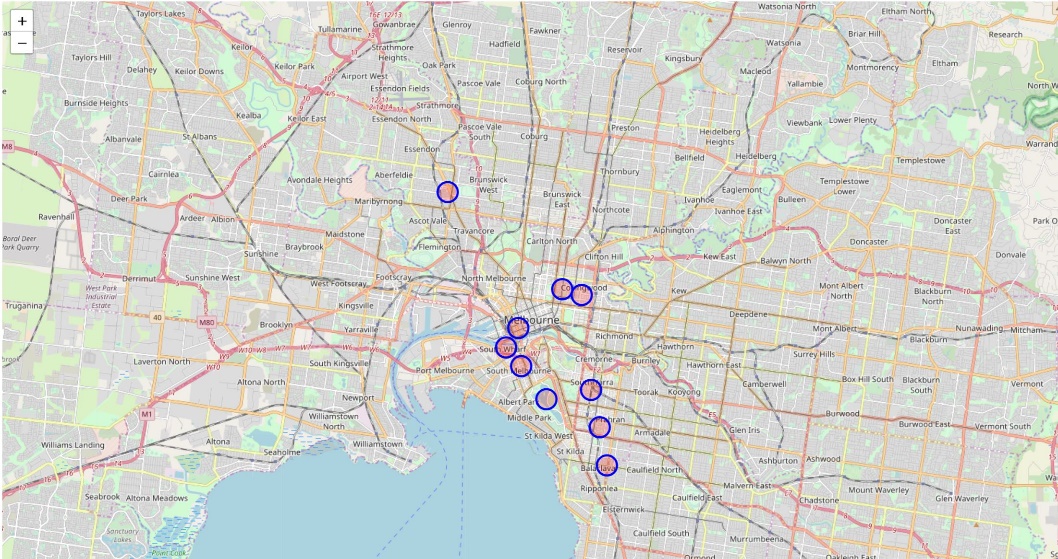
It can be seen that most venues are densely located in and around the central business district of Melbourne. One of the clusters (coloured in orange) only contains a single location, on the south end of Footscray - west of the city centre. This outlier should be taken into consideration when determining the ideal locations of the shops. The means of the clusters were also visualised:



## 10 Suburbs with the most gyms



The 10 suburbs which contained the highest number of gyms were found. The suburb of Melbourne (which encompasses the inner-city business district) had the most with 14, and South Melbourne was second with 6. It is important to note that there are many more suburbs apart from Fitzroy and ‘Chapel Street North’ that only contain 3 venues, and therefore should not be considered as they only on the list due to chance.

The map below visualises each of the suburbs:

# Discussion

The main finding of the analysis was that Melbourne’s central business district contains the highest number of venues and would therefore act as an ideal location for the company’s flagship store. This finding is supported by the k-means analysis, as two cluster means are located on each side of the CDB. However, the validity of the results may have been affected by numerous extraneous variables, such as the fact that different gyms tailor to different age-groups and will therefore differ in the number of members who will purchase supplements. Despite this, it can be said relatively conclusively that Melbourne’s inner city is a prime location, as it contains more than twice as many venues than any other suburb.

Furthermore, the k-means analysis revealed roughly 6 distinct clusters of venues, the means of which are located in Prahran, Melbourne CBD, North Melbourne, Collingwood, Hawthorn and Ascot Vale – all of which would undoubtedly be profitable locations. This is reinforced by the correlation between the results of both techniques. Other cluster means, such as the one in Footscray, should however not be prioritised as locations due to the low number of venues in the clusters.

# Conclusion

Ultimately, it can be concluded that suburbs in and around the centre of Melbourne (as well as Ascot Vale), would be suitable locations for ‘Bulk like Hulk’ stores. Contrary to stakeholder’s demands, the idea to not locate two venues in the same suburb should be reconsidered, as the Melbourne business district has a much higher density of venues than any other suburb by a substantial margin (it is, however, recommended to analyse the presence of competition).