**Capstone Project – Battle of Neighborhoods**

Contents

[Chapter 1 : Introduction 1-2](#_Toc34055036)

[1.1 Background 1-2](#_Toc34055037)

[1.2 Problem Description 1-2](#_Toc34055038)

[1.3 Project Objectives 1-3](#_Toc34055039)

[1.4 Project Scope 1-3](#_Toc34055040)

[Chapter 2 : Data 2-4](#_Toc34055041)

[2.1 Introduction 2-4](#_Toc34055042)

[2.2 Data Requirements 2-4](#_Toc34055043)

[2.3 Source of Data 2-5](#_Toc34055044)

[2.4 Validity / Reliability of Data 2-6](#_Toc34055045)

[Chapter 3 : Methodology 3-7](#_Toc34055046)

[3.1 Introduction 3-7](#_Toc34055047)

[3.2 Design of the methodology 3-7](#_Toc34055048)

[3.3 Initial data analysis 3-7](#_Toc34055049)

[3.4 Exploratory data analysis 3-8](#_Toc34055050)

[3.4.1 Clustering………………………………………………………….3-8](#_Toc34055051)

[3.4.2 Machine Learning used for Clustering – K-Means………………3-9](#_Toc34055052)

[3.4.3 Evaluation – Elbow method……………………………………….3-9](#_Toc34055053)

[Chapter 4 : Results 4-10](#_Toc34055054)

[4.1 Our findings: 4-10](#_Toc34055055)

[4.2 Shortcomings: 4-10](#_Toc34055056)

[Chapter 5 : Discussion 5-11](#_Toc34055057)

[5.1 Observations 5-11](#_Toc34055058)

[5.1.1 Data……………………………………………………………….5-11](#_Toc34055059)

[5.1.2 Methodology………………………………………………………5-11](#_Toc34055060)

[5.2 Recommendations 5-11](#_Toc34055061)

[Chapter 6 : Conclusion 6-12](#_Toc34055062)

# Introduction

## Background

Our client, Mrs. Grace Smith is a retired school teacher who lives in a suburb named Docklands in Melbourne. In Australian English the term ‘suburb’ is largely synonymous with what is called a ‘neighborhood’ in other countries. Now, due to some personal reasons, Mrs. Smith needs to move to another suburb in Melbourne. She finds Docklands very comfortable, so she wishes to move to a neighborhood that is very similar to Docklands.

## Problem Description

In order to proceed further, we needed to understand the problem more specifically. Hence we needed precise answers to the following questions.

* Is Mrs. Smith okay with moving out of Melbourne, i.e. to a neighborhood in any other nearby city in Australia?
* What did Mrs. Smith mean by ‘very similar’ to Docklands?

So, we conducted interviews with Mrs. Smith to understand her lifestyle and preferences. We understood that she is 61 years old and lives with her husband, she leads an active life, has friends mostly living in Melbourne, goes for a walk in the park every morning and evening, buys groceries herself , uses public transport, prefers to eat home cooked food and so on….

From these interviews, we gathered that

* Mrs. Smith does not wish to move outside Melbourne city.
* By ‘very similar’ to Docklands, she meant – the presence of venues like parks, grocery stores, supermarkets, good transportation facilities like car rentals, light rail stations, etc. She was not particular about venues like cafes, restaurants, schools or colleges.

## Project Objectives

Make a recommendation to our client, Mrs. Smith on the following :

* Choice of a suburb (neighborhood) in Melbourne that is very similar to Docklands.
* The similarity should be based on the presence of particular venues that the client mentioned/needs to ensure that she can live comfortably.

## Project Scope

Neighborhoods could be compared based on several different parameters including cost of living, demographics, culture, presence of schools, emergency services, law and order situation, noise, etc. But these parameters are outside the scope of this project.

This project aims at recommending a neighborhood that is very similar to Docklands as per the needs of Mrs. Smith only.

# Data

## Introduction

Here, we shall look into the data that will be required to solve our problem and the source of the data. We shall also ensure the validity and reliability of this data.

## Data Requirements

Our problem is about comparing neighborhoods in a city. So we will need to leverage data from a location data provider to know about which neighborhoods are present in Melbourne city and about the venues in each of these neighborhoods. Location data platforms mostly work based on geographic coordinate data, which we would need to provide.

So, we could list our data requirements as follows:

* Names of all neighborhoods in Melbourne city.

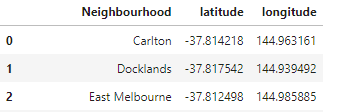
Example data include : , Parkville, East Melbourne

* List of venue categories that are preferred by Mrs. Smith

Example venue categories include: Supermarket, Park, Light Rail Station, Hospital

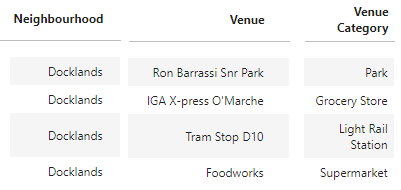
* Geographic coordinates (Latitude and Longitude) of Melbourne city and all the neighborhoods in Melbourne city.

Example:



* Venues present in each of the neighborhoods that match Mrs.Smith’s preferences

Example:



## Source of Data

* **Names of all neighborhoods in Melbourne city.**

The website - <https://www.melbourne.vic.gov.au/about-melbourne/melbourne-profile/suburbs/Pages/suburbs.aspx> is a government website which provides information about all the suburbs in Melbourne city. The required data needs to be extracted using web-scrapping.

* **Venues present in each of these neighborhoods.**

Location data provider called Foursquare provides up-to-date information about venues.

* **List of venue categories that are preferred by Mrs. Smith**

Firstly, we would collect the venue categories that Mrs. Smith is looking for in the neighborhood. These would then be matched with Foursquare’s Venue Categories. The ‘Developer’ section of Foursquare provides a list of all Venue Categories used in Foursquare at <https://developer.foursquare.com/docs/resources/categories>.

* **Geographic coordinates (Latitude and Longitude) information of Melbourne city and all the neighborhoods in Melbourne city.**

At the moment the plan is to use Geocoder to generate this data.

## Validity / Reliability of Data

The data in Foursquare is believed to be reliable based on geospatial coverage, accuracy and update frequency.

The initial plan was to extract the names of neighborhoods in Melbourne city from the Corra website-<http://www.corra.com.au/australian-postcode-location-data/>. This was an attractive option because it provided neighborhood data along with the latitude and longitude information in a csv file format. This meant that extracting data would be a simple straight-forward process. But since the data was last updated on March 19th, 2013 and the site itself claimed, “ PLEASE NOTE: THIS DATA IS NOW OUT OF DATE BY QUITE A BIT”, the data didn’t seem fit for our project. So the Australian government website was chosen over this one, even though extracting data would need web-scrapping.

We will be using Geocoder to generate data on geographic coordinates. A problem with Geocoder is that sometimes it fails to generate data for some neighborhoods. Incase this happens, we would have to look for other options. But that shouldn’t be a problem.

# Methodology

## Introduction

In this chapter, we shall look into how the initial data analysis was performed, the exploratory data analysis, the machine learnings that are planned and the reasons for choosing these methods. We shall also look into the evaluation methods used.

## Design of the methodology

The figure below depicts the steps that we plan to use in this project.

Figure 1- Design of the methodology

## Initial data analysis

During the initial data analysis, we refrained from any analysis that was aimed at answering the original research questions. This step was more about checking the validity and reliability of the data. We also looked for missing data and planned ways to deal with such data. The outliers were also removed. We mostly used cross-validation techniques here.

## 

## Exploratory data analysis

Here, our intention is to find one or more neighborhoods that are very similar to Docklands. That is, we need to *group* similar neighborhoods in Docklands. But we do not have any reference data or training data. So using Classification, which deals with supervised, labelled data for exploratory data analysis is out of question. Hence the option available to us for grouping the neighborhoods is – ‘Clustering’ because it works on unsupervised, unlabeled data.

### 3.4.1 Clustering

**Clustering** is a common exploratory data analysis technique which groups a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).

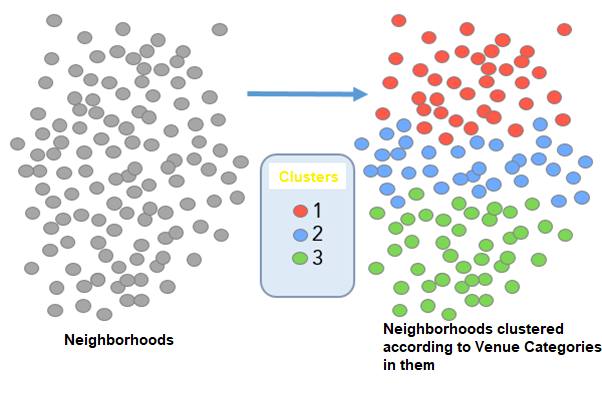


Figure 2 - Clustering

So, using this technique we can group neighborhoods which have similar venues together into clusters. Then we can further study the neighborhoods in the cluster that Docklands belongs to.

### Machine Learning used for Clustering – K-Means

Next, we have to plan about which Machine Learning Clustering algorithm suits our needs. For that, lets consider our data and requirements. We do not have outliers in our data, so there is no need to use a complex algorithm that deals with noise. We also do not need to create a hierarchy for clustering.  So hierarchical clustering also is not required. K-Means seems easy and fast to implement as all we’re really doing in K-Means is computing the distances between points and group centers. So K-Means satisfies our needs.

### Evaluation – ****Elbow**** method

K-Means requires the number of clusters (*K)*as an input and doesn’t learn it from data. There is no right answer for K that we should have in any problem. So we will use the Elbow method which gives an idea on what a good *K* would be, based on the sum of squared distance (SSE) between data points and their assigned clusters’ centroids. As shown in the figure below, we pick *k* at the spot where SSE starts to flatten out and forming an elbow.

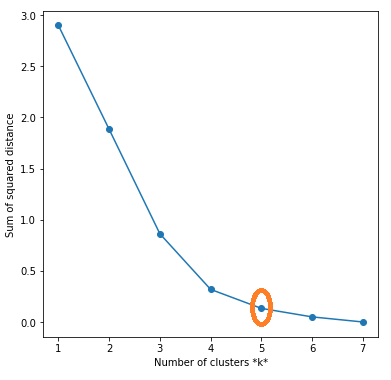


Figure 3-Elbow Method

# Results

The objective of this project was to make a recommendation to our client, Mrs. Smith on the following:

* Choice of a suburb (neighborhood) in Melbourne that is very similar to Docklands.
* The similarity was based on the presence of particular venues that she mentioned/needed to ensure that she could live comfortably.

## Our findings:

‘Carlton’ is the suburb that is most similar to Docklands as shown in the map below (marked in red)

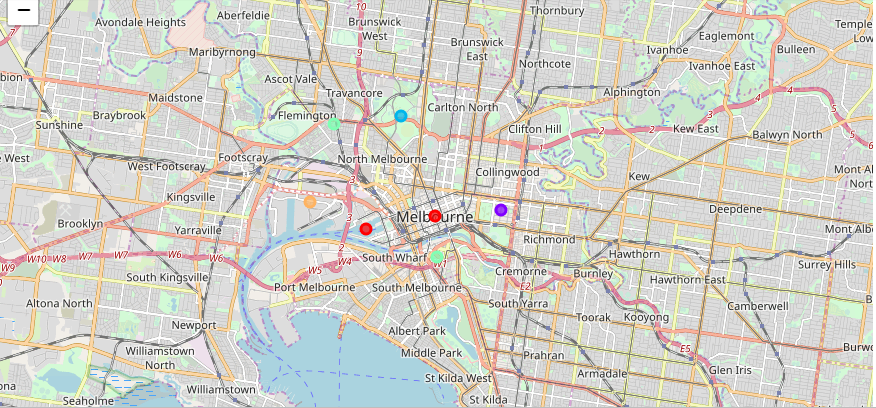


Figure - Melbourne Map after clustering

* The following venue categories were found that Mrs. Smith needed:

Park, Grocery Store, Beer Garden, Flea Market, Light Rail Station, Train Station,

Sculpture Garden, Supermarket.

* The bar chart below shows how many venues of each the Venue Category were found in Carton

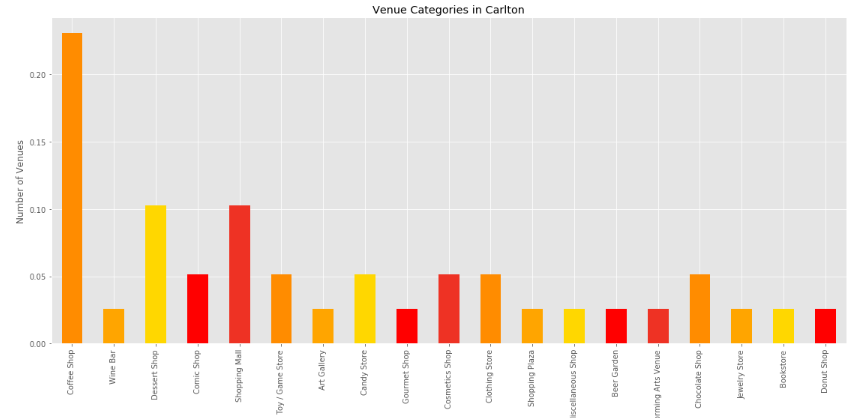


Figure - Bar chart of Venues in Melbourne

## Shortcomings:

The study could not find some venue categories like Car rental, Bus station, Hospital. We used a web search and manual methods to search for the availability of the missing venue categories. For example, we found the following venues -

|  |  |  |
| --- | --- | --- |
| **Neighbourhood** | **Venue** | **Venue Category** |
| Kensington and Flemington | The Royal Women's Hospital | Hospital |
| Kensington and Flemington | Guardian Medical Centre Flemington | Health Care Centre |
| Kensington and Flemington | Kensington Station | Bus stop |
| Kensington and Flemington | Flemington Racecourse Station | Train station |
| Kensington and Flemington | GoGet CarShare Pod | Car rentals |

# Discussion

## Observations

During the course of this project, certain observations were made regarding the data, the methodology adopted and the result.

### Data

It was observed that Foursquare mostly provided location data about social venues like cafes and restaurants than venues like bus stations or car rentals. Perhaps this is because Foursquare data is community-sourced and most people generally do not offer reviews about such venues.

### Methodology

We used Unsupervised Machine Learning here because of the unavailability of labelled data. But the methodology had its disadvantages as well. It was computationally more complex than Supervised methods. Evaluating the accuracy of the results was not straight forward as with unsupervised learning generally. We did not have any ground truth to compare with or a solid evaluation metric. The Elbow method did help with the evaluation process though.

### Result

We wished to ensure that Carlton was not only most similar to Docklands, but that it was also the neighborhood that suited Mrs.Smith’s needs the most. A Word Cloud was used to visualize this. It showed that Carlton was indeed the town that had most required venues. In fact Carlton was better than Docklands too.



Figure - Word Cloud

## 

## 5.2 Recommendations

Clustering as a machine learning technique might not give us very accurate results and evaluation is also difficult. But it can be effectively used in these scenarios:

* + When there is a large amount of unlabeled data, grouping the data into clusters can give us meaningful intuition of the structure of the data we’re dealing with.
  + When data seems to have many sub-groups, each of which seem very different in their behaviors, the data can be clustered first and then models could be built for each of these clusters to be used as training data. For example, for this project, the initial plan was to cluster the neighborhoods and then prepare a decision tree model which would enable Mrs. Smith to choose the required neighborhood. (But Melbourne is a small city, so clustering itself gave us an insight into the result

# Conclusion

We studied Melbourne city and found that Carlton is the neighborhood that is very similar to Docklands. Labelled data was not available, so unsupervised machine learning was employed. Elbow method, domain knowledge and intuition helped in evaluating the result. We observed that the result of Machine Learning is largely dependent on the quality of data. Getting data that is accurate, complete, valid, reliable, timely, relevant and accessible is challenging. Also, the result depends on our ability to choose the right algorithm and to interpret results generated by the algorithms. But yes, Machine Learning if applied correctly and effectively does hold the potential to revolutionize the way things work.

**References**

Chris Albon(2014). Group Bar Plot In MatPlotLib[Online]. Available fromhttps://chrisalbon.com/python/data\_visualization/matplotlib\_grouped\_bar\_plot/[accessed 03-March-2015]

# Exploratory Data Analysis[Online]. Available from <https://en.wikipedia.org/wiki/Exploratory_data_analysis>[accessed 01-March-2015]

# [George Seif](https://towardsdatascience.com/@george.seif94?source=post_page-----a36d136ef68----------------------)(2018). The 5 Clustering Algorithms Data Scientists Need to Know. [Online]. Available from <https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68> [accessed 01-March-2015]

# [Imad Dabbura](https://towardsdatascience.com/@ImadPhd?source=post_page-----aa03e644b48a----------------------)(2018). K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. [Online]. Available from<https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a> [accessed 01-March-2015]