**Main Objective of the Analysis:**

The main goal of this project is to collect and analyze data in order to select a location in Melbourne to open a Cafeteria. We want to help a business owner planning to open up a Cafe in a location by exploring better facilities around the Suburb.

This is an unsupervised machine learning problem where we need to group together

suburbs having similar facilities. We will use K Means Clustering and Hierarchical Clustering to solve this problem.

**Data Description:**

- List of Suburbs in Melbourne, Australia which I have extracted from: https://en.wikipedia.org/wiki/Category:Suburbs\_of\_Melbourne

- Latitude & Longitude of all the suburbs using Geocoder- venues in each suburb from foursquare API https://foursquare.com/





**Data Understanding**

- The Wikipedia page contains a list of suburbs in Melbourne. There are 212 suburbs in Melbourne which I extracted using a web scraping technique with the help of Python BeautifulSoup and Request packages.

- the geographical coordinates such as latitude and longitude of each suburb were

collected using Python’s Geocoder package.

- Then, Foursquare API was used to extract details about the various venues present

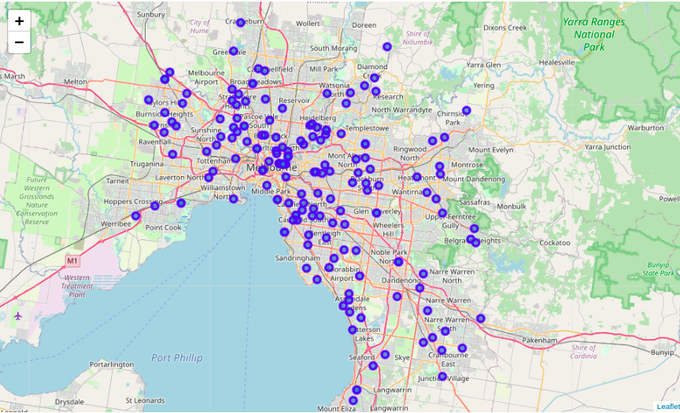
in each suburb.

- Once, the location data was extracted by using Geocoder, I used the Folium package

to visualize the data on a map. This ensured us that the data we retrieved was correct.

- Foursquare API was used to obtain the top 100 venues within a radius of 2000

meters.



**Data Cleaning and Feature Engineering**

- Converted the data into dummy variables using get\_dummies method of Pandas

package that will be essential for performing clustering algorithm

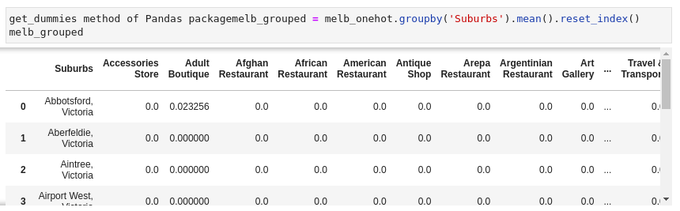
- Grouped the data by Suburb & also taking the mean of the frequency of occurrence

of each category.

- I extracted the data of the Cafeteria only

- Our final data frame had two variables: suburb name and the mean of the frequency

of occurrence of cafes





**Modeling**

- Performed clustering on the data using K-means clustering and Hierarchical Clustering.

- For K means Clustering I used k = 3, 4, 5 clusters based on the frequency of occurrence of Cafes in each suburb.

- Found out the suburb which had the highest concentration of Cafes and also the

lowest concentration

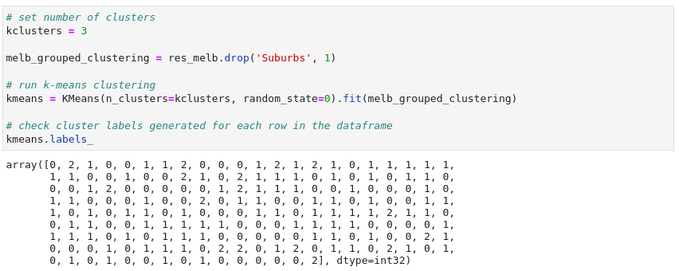
**Results**

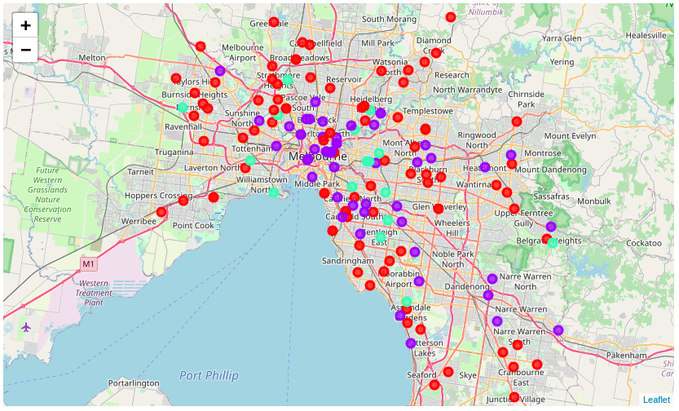
I decided to use 3 clusters for this problem as this gives the best result. Categorized the data into 3 categories using K-means clustering based on the frequency of occurrence for ‘Cafe’.

- Cluster 0: Suburbs with a low number of Cafes.

- Cluster 1: Suburbs with a moderate number of cafes.

- Cluster 2: Suburbs with a high concentration of Cafe.





**Evaluation**

- Cluster 0 is displayed as the red color represents a greater opportunity and high

potential but also suffers from the risk of having fewer customers as those areas are

not busy areas.

- As a new business owner it wouldn’t be wise enough to choose cluster 2. Therefore, I would recommend that cluster 1 represented by blue color, should be chosen where there is medium competition but greater opportunity.



**Suggestions for Next Step:**

I could get the population and average income of the suburbs, and then calculate the

money to cafeterias ratio = population \* income / number of cafes. The suburb with

highest ratio would be the best opportunities, as they have a lot of population and

money but less competition.