Classifying two major cities in Australia

Introduction

Main idea of this project is to help new immigrant to identify best suburbs for settlement where preferred facilities are available. This Project would help the immigrant take a better decision on choosing the best neighborhood out of many suburbs to rent their houses in Sydney/Melbourne based on the distribution of various facilities in and around that neighborhood. As an example, this project would compare all the suburbs neighborhoods and analyse the distribution of facility.

Also, this project uses K-mean clustering unsupervised machine learning algorithm to cluster the venues based on the place category such as Train/Bus stop, Shopping mall, movie, gym etc. This would give a better understanding of the similarities and dissimilarities between the suburbs neighborhoods to retrieve more insights.

Immigrants are much more likely to settle in capital cities, especially in inner city suburbs or suburbs near universities, than the Australian-born population. Settlement patterns also vary by visa type and country of origin. (Source:

https://www.pc.gov.au/inquiries/completed/migrant-intake/report/migrant-intake-report.pdf)

Background

New South Wales (NSW): New South Wales (abbreviated as NSW) is a state on the east coast of Australia. It borders Queensland to the north, Victoria to the south, and South Australia to the west. Its coast borders the Tasman Sea to the east. The Australian Capital Territory is an enclave within the state. New South Wales' state capital is Sydney, which is also Australia's most populous city. In March 2018, the

population of New South Wales was over 7.9 million,[3] making it Australia's most populous state. Just under two-thirds of the state's population, 5.1 million, live in the Greater Sydney area.[9] Inhabitants of New South Wales are referred to as New South Welshmen.[1][2] (Source: https://en.wikipedia.org/wiki/New_South_Wales)

Victoria (VIC): Victoria (abbreviated as Vic) is a state in south-eastern Australia. Victoria is Australia's most densely populated state and its second-most populous state overall. Most of its population lives concentrated in the area surrounding Port Phillip Bay, which includes the metropolitan area of its state capital and largest city, Melbourne, Australia's second-largest city. Geographically the smallest state on the Australian mainland, Victoria is bordered by Bass Strait and Tasmania to the south, [note 1] New South Wales to the north, the Tasman Sea to the east, and South Australia to the west. (Source: https://en.wikipedia.org/wiki/Victoria_(Australia))

Immigrants are more likely to settle in urban areas than people born in Australia

Objective

The objective of this project is to find best suitable suburb in Sydney and Melbourne given that new settler have some priority of facility in the suburb. using suburb location data along with Foursquare data and machine learning segmentation and clustering, this project will recommend cluster of suburbs where new settler will be able to get preferred facility.

1. Movie 2. Shopping Mall 3. Turkish Restaurant 4. Bus 5. Train 6. Fish 7. Gym

Aim of this project:

1) Classify Sydney suburbs' based on given preferences and find best suburb that will meet requirements

2) Reclassify Sydney suburb given that house rent, travel time and distance will be minimized, and also meet requirements as per preference and weight

3) Predict Sydney house rent for given facility requirement

4) Classify Melbourne suburbs' based on given preferences and find best suburb that will meet requirements

Target Audience

Through this project we are expecting following people to benefit out of the findings.
- Tourist.
- People migrating city for work.
- Business person looking for new location to start office etc.
- And many more.

Tools and Data

Australia Postcode data, suburb data with lat long, [this data added to github as csv file]

- Foursquare data
- Foursquare API
- IBMWatson Account
- Jupiter Notebooks
- Synthetic Data [program code included]
- Python packages and Dependencies:

• Pandas - Library for Data Analysis

• NumPy – Library to handle data in a vectorized manner

• JSON – Library to handle JSON files

• Geopy – To retrieve Location Data

• Requests – Library to handle http requests

• Matplotlib – Python Plotting Module

• Sklearn – Python machine learning Library

• Folium – Map rendering Library

Preferred Weight

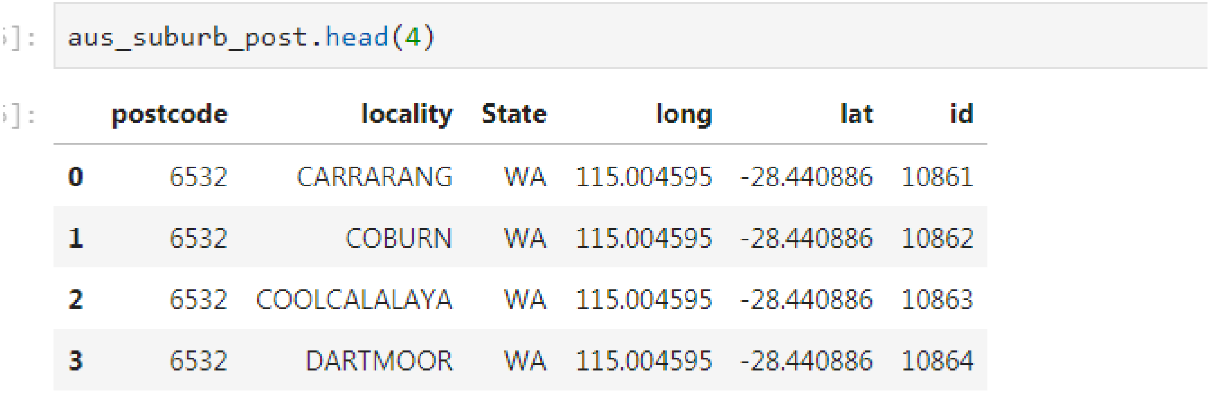
- Train Station: 2
- Shopping Mall: 1.8
- Bus Station:1. 5
- Turkish Restaurant: 1.4
- Gym: 1.3
- Movie: 1.0
- Fish: 1.0

Synthetic data

- House rent
- Travel time
- Distance of suburb from Sydney city centre

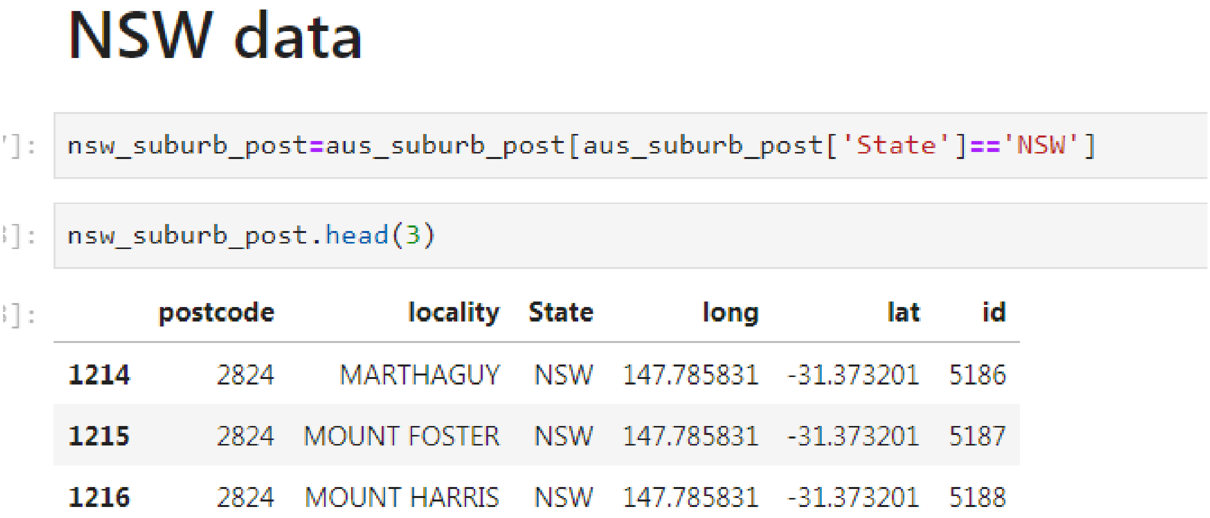
Data needed:

We will use geolocation data dataset- Australia Post code data set, where post code, suburb name and lat long given. This data will be uploaded to github. This data will be used to identifying Sydney and Melbourne Suburbs.



```
aus_suburb_post.head(4)
```

	postcode	locality	State	long	lat	id
0	6532	CARRARANG	WA	115.004595	-28.440886	10861
1	6532	COBURN	WA	115.004595	-28.440886	10862
2	6532	COOLCALALAYA	WA	115.004595	-28.440886	10863
3	6532	DARTMOOR	WA	115.004595	-28.440886	10864



NSW data

```
nsw_suburb_post=aus_suburb_post[aus_suburb_post['State']=='NSW']
```

```
nsw_suburb_post.head(3)
```

	postcode	locality	State	long	lat	id
1214	2824	MARTHAGUY	NSW	147.785831	-31.373201	5186
1215	2824	MOUNT FOSTER	NSW	147.785831	-31.373201	5187
1216	2824	MOUNT HARRIS	NSW	147.785831	-31.373201	5188

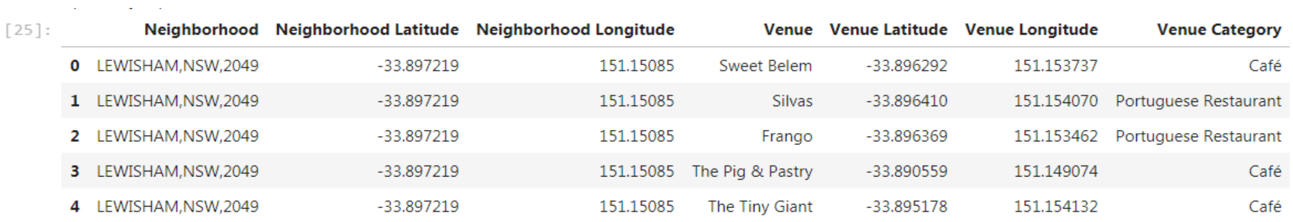
Foursquare data

This data will be collected from Foursquare web site using API.

Foursquare API will be used to query each of the neighbourhood

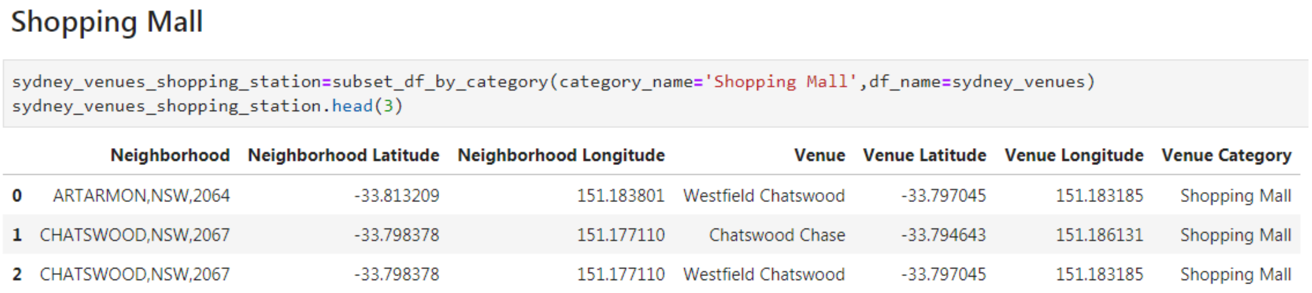
Data frame will be created for preferred facilities

(1. Movie 2. Shopping Mall 3.Turkish Restaurant 4. Bus 5. Train 6. Fish 7. Gym)



[25]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	LEWISHAM,NSW,2049	-33.897219	151.15085	Sweet Belem	-33.896292	151.153737	Café
1	LEWISHAM,NSW,2049	-33.897219	151.15085	Silvas	-33.896410	151.154070	Portuguese Restaurant
2	LEWISHAM,NSW,2049	-33.897219	151.15085	Frango	-33.896369	151.153462	Portuguese Restaurant
3	LEWISHAM,NSW,2049	-33.897219	151.15085	The Pig & Pastry	-33.890559	151.149074	Café
4	LEWISHAM,NSW,2049	-33.897219	151.15085	The Tiny Giant	-33.895178	151.154132	Café



Shopping Mall

```
sydney_venues_shopping_station=subset_df_by_category(category_name='Shopping Mall',df_name=sydney_venues)
sydney_venues_shopping_station.head(3)
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	ARTARMON,NSW,2064	-33.813209	151.183801	Westfield Chatswood	-33.797045	151.183185	Shopping Mall
1	CHATSWOOD,NSW,2067	-33.798378	151.177110	Chatswood Chase	-33.794643	151.186131	Shopping Mall
2	CHATSWOOD,NSW,2067	-33.798378	151.177110	Westfield Chatswood	-33.797045	151.183185	Shopping Mall

Methodology

In order to proceed with this research, we need to read Australia postcode data into pandas data frame. Then need to find suburbs nearby Sydney and Melbourne. In this project suburbs which are located far from city area excluded from the analysis. Therefore, for Sydney, NSW postcodes from NSW 2000 to NSW 2100 and for Melbourne VIC 3000 to VIC 3100, suburbs are considered.

After short listing Sydney suburbs, the Foursquare API was then used to query each of the neighborhood. From the resulting neighborhood data, data frame for preferred facilities (1. Movie 2. Shopping Mall 3.Turkish Restaurant 4. Bus 5. Train 6. Fish 7. Gym) are extracted and new data frame is created.

Next task is to calculate points based on the client weight preference.

1. Movie 2. Shopping Mall 3.Turkish Restaurant 4. Bus 5. Train 6. Fish 7. Gym

If suburb contain 1 venue then points is the weight multiplied by 10, but if the suburb contain more than one venue, then that suburb will get additional 1 points for each additional venue present.

venue points= 10*weight + (number of venue -1)

total points= sum of venue points

Total number of venue counts is based on the total types of facility/client requirement available at the suburb. Total venue points ranges from 0 to 7. If any suburb contains all 7 client requirements (1. Movie 2. Shopping Mall 3.Turkish Restaurant 4. Bus 5. Train 6. Fish 7. Gym) then the suburb will be awarded 7.0 points.

Unsupervised machine learning algorithm K-mean clustering would be applied to form the clusters of different categories of places residing in and around the neighborhoods. The resulting data frame will be used for Cluster analysis. For cluster analysis, Total points and number of venue used. After cluster analysis, cluster centres were plotted on a Folium map.

This methodology is used for Sydney and Melbourne suburbs. From the resulting data frame, cluster of suburbs with highest number of points, venues extracted and presented.

For Sydney area, I have included Synthetic data of House rent, Travel time and Distance from the Sydney city centre. Objective of this part is to find best suburbs if client would like to know whether previous selection will remain same or name.

I have calculated distance using haversine distance function, then synthetically generated rent and travel time using function. For this three attributes, weight is assumed for point calculation.

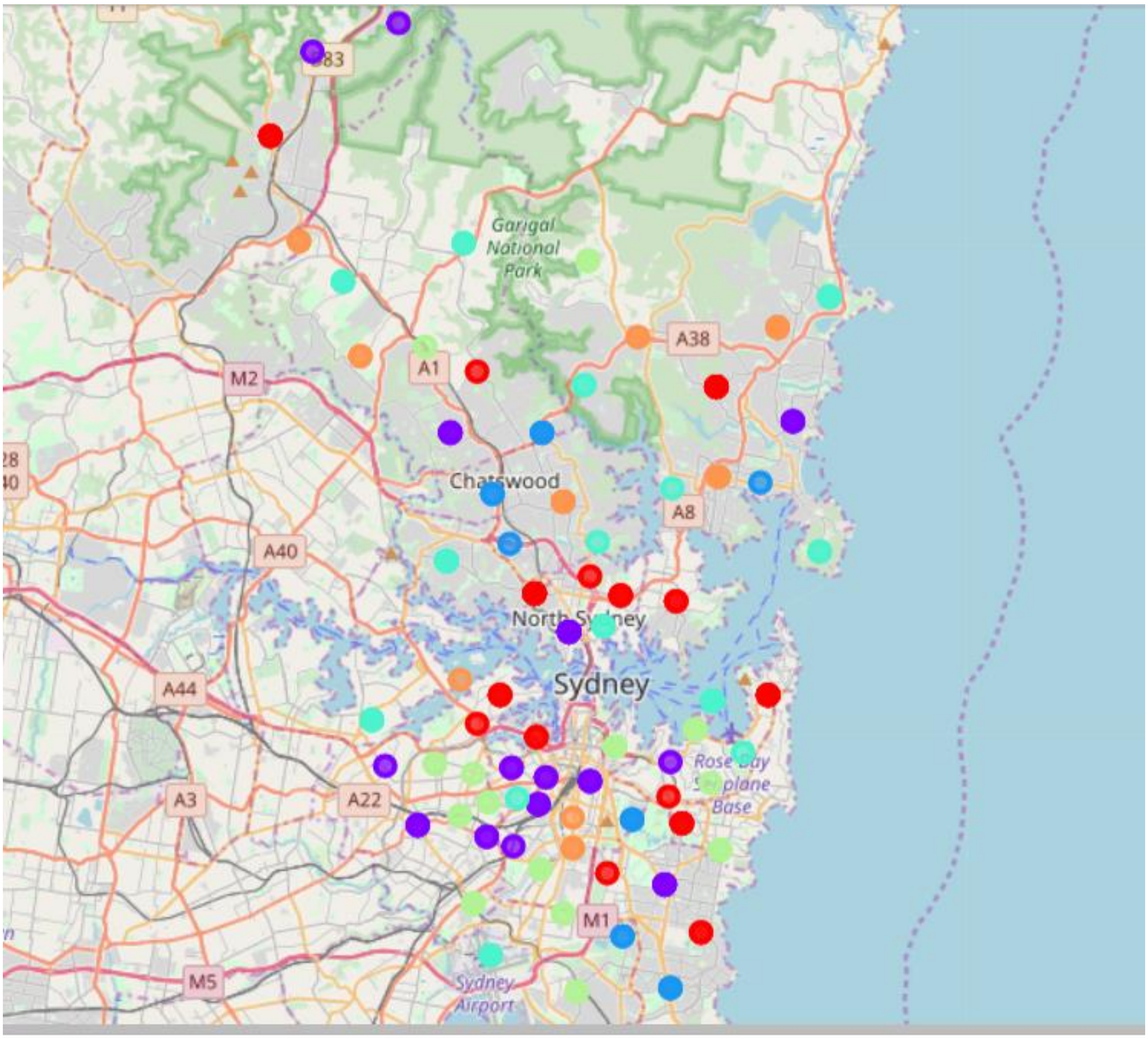
After adding this feature, i have re-calculated total points and saved data into dataframe. Finally, this new total points is used for cluster analysis.

In addition, predictive for Sydney house rent is developed given the client facility requirement. To do this task, I used KNN algorithm and the Sydney dataframe, which is used for clustering.

Results

Considering availability of Train Station, Bus Station, Fish, Gym, Movie and Turkish Restaurant within the 2 km distance of suburb

list of best Sydney Suburb given below:



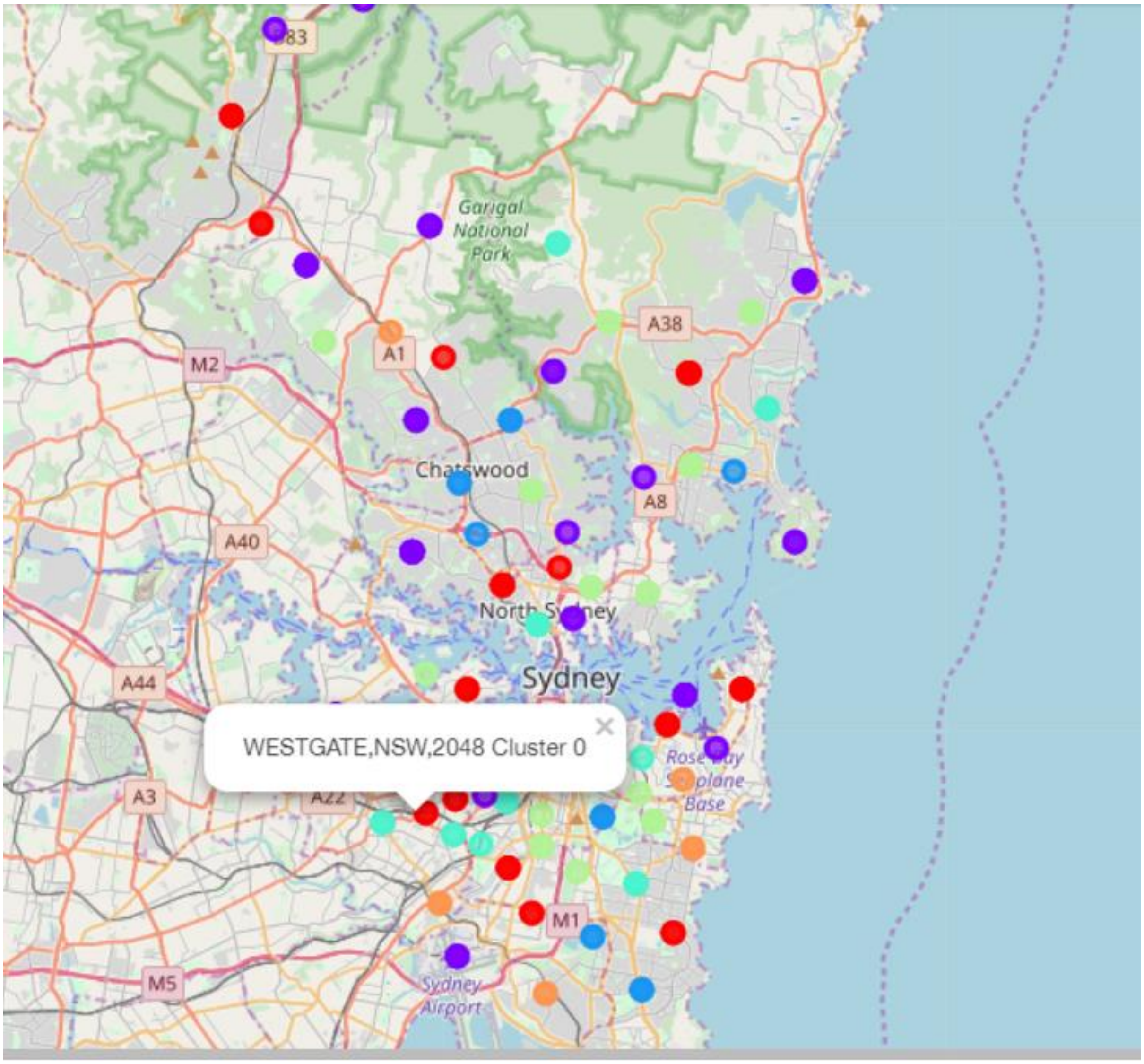
'CASTLE COVE,NSW,2069', 'ROSEVILLE,NSW,2069', 'ROSEVILLE

CHASE,NSW,2069','PAGEWOOD,NSW,2035','MAROUBRA,NSW,2035', 'MAROUBRA SOUTH,NSW,2035','KINGSFORD,NSW,2032', 'CENTENNIAL

PARK,NSW,2021','DACEYVILLE,NSW,2032','PADDINGTON,NSW,2021''

If some other factors such as Rent, Travel_time, Distance_CBD included in the decision making process then some of the suburb previously selected may not be good option.

list of best Sydney Suburb given below:



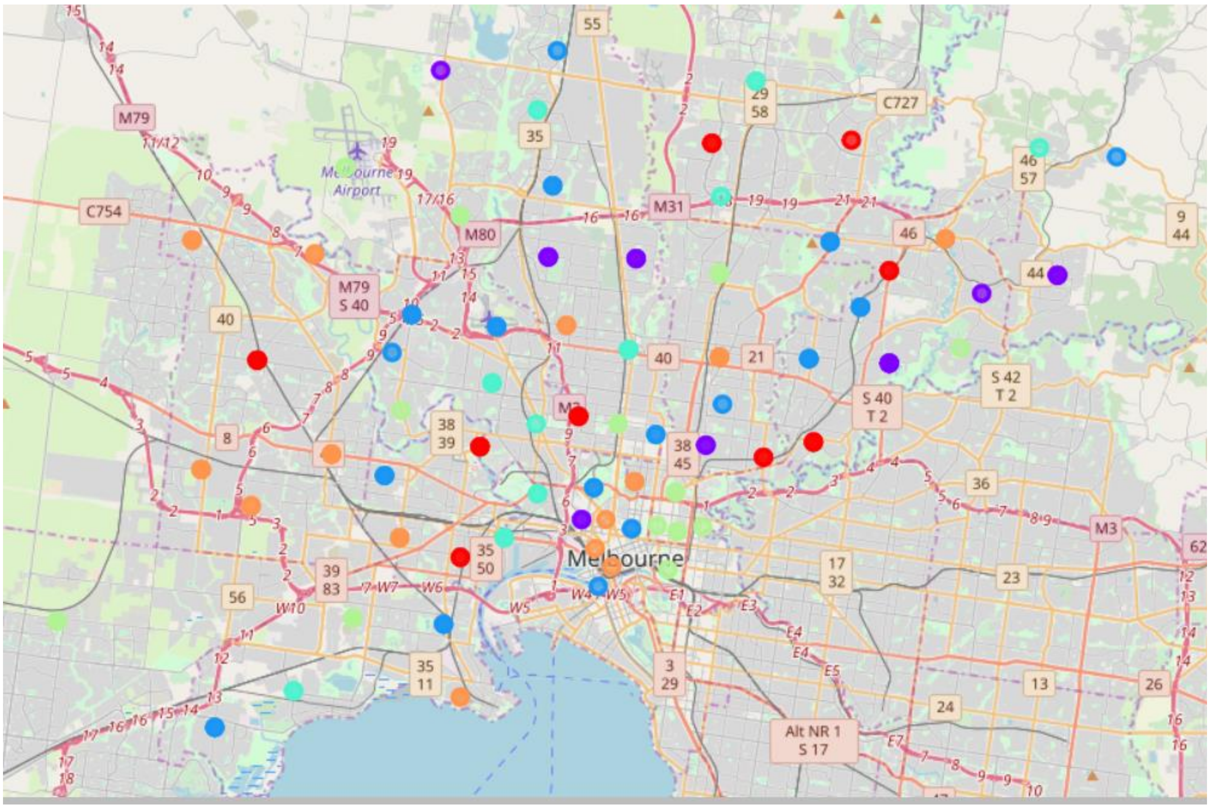
'PAGEWOOD,NSW,2035', 'MAROUBRA,NSW,2035','MAROUBRA

SOUTH,NSW,2035','CENTENNIAL PARK,NSW,2021','PADDINGTON,NSW,2021','MOORE

PARK,NSW,2021','ROSEVILLE,NSW,2069','ROSEVILLE CHASE,NSW,2069','CASTLE

COVE,NSW,2069', 'KINGSFORD,NSW,2032'

list of best Melbourne Suburb given below:



EPPING DC,VIC,3076 EPPING,VIC,3076 ABERFELDIE,VIC,3040 ESSENDON,VIC,3040 ESSENDON WEST,VIC,3040 KENSINGTON,VIC,3031 FLEMINGTON,VIC,3031 THOMASTOWN,VIC,3074 MOONEE PONDS,VIC,3039 COOLAROO,VIC,3048

Conclusions

New immigrant or any person moving can use this project to identify the suitable place depending on the individual preferences of facilities. This project will also help to refine decision my adding other constraints such as weekly rent budget, travel distance and time. It is clear from the analysis that selection best suitable suburbs may change based on the preference and restrictions. Here also developed model using KNN to predict house rent, using another input variable.