Analytical view of City of Melbourne and City

Introduction

Background: Melbourne and Sydney are the 2 main cities when living in Australia is considered or analysed. Although there are other beautiful and growing cities in Australia,People outside of Melbourne mainly look at the living standards and day to day life of these 2 cities mainly. Both the cities are similar in natural beauty, resource, places to visit, population, diversity, job opportunity and career opportunities. Yet an analytical view comparing the cities will have a very vast user base. My report will mainly target the migrant population to Australia

Problem: Australia has a very good migration policy and every year Australia accepts nearly 200,000 population as part of skilled migration programs. These people migrating to Australia normally do not have an analytical view of which City they have to land after receiving their permanent residency visa.Normally, these people talk to friends , colleagues and online forums to create an understanding of cities of Australia.

Interest: My analysis will use clustering methods to cluster the suburbs of both the city and display various characteristics of the suburbs in these cities. Readers will be able view data for schools, restaurants, playgrounds, places to visit and jobs created for major suburbs in these cities and create a perspective of both the city on their own. This analytical view will not give an idea of which city is better rather this report will only provide a visual analysis of the major suburbs of the city and leave it to the reader to create a perspective.

Data Acquisition

I would be mostly working on the below dataset and as I start developing all the analytical view would add any other required data:
Suburb Coordinates: I am planning to do some web scraping to find the coordinates for Suburbs of Melbourne and use the kaggle dataset for Sydney suburbs : Sydney Suburb Coordinates.
Suburb Characteristics: Suburbs and it's nearby venue information will be fetched from foursquare places API.
Melbourne Housing Data: Use the Kaggle dataset for Melbourne housing data: Melbourne_housing_data
Sydney housing data : Extract house pricing dataset from NSW Government Sales Dataset
Extract **postcode of Sydney and Melbourne** from Australia post API: Search Suburb/postcode API

Suburb and its coordinate data will be used to segregate each suburb and put them on the map. Suburb Characteristics will help provide comparison analysis between Sydney and Melbourne.

Data Wrangling

As part of Data wrangling Process I would be extracting following set of Data:

1. Suburb, Co-ordinates , postcode and Locality
2. Housing Sales data
3. Suburb Features data

Prepare data for Melbourne: Melbourne Housing data has the required suburb, postcode and Co-ordinates so I extracted the required data and cleaned the data into the required format and feature.

	Suburb	Postcode	Regionname	Lattitude	Longtitude
0	Abbotsford	3067.0	Northern Metropolitan	-37.8014	144.9958
66	Airport West	3042.0	Western Metropolitan	-37.7180	144.8780
133	Albert Park	3206.0	Southern Metropolitan	-37.8459	144.9555
194	Alphington	3078.0	Northern Metropolitan	-37.7818	145.0198
230	Altona	3018.0	Western Metropolitan	-37.8700	144.8250

Now, there are a lot of suburbs in Melbourne and they are divided into different regions. As part of my analysis, I would be restricting this list only to Eastern Metropolitan region. The reason for doing so is to limit the data which I will be fetching from foursquare as there is a cap on the number of API calls from Foursquare. Also, It gives the analysis of displaying data only for a part of the city. People normally select the region first before selecting the suburb when they plan to live in a suburb. This solution can be extended to let users choose a particular region from a UI field and run the rest of the analysis based on that.

:

	Suburb	Postcode	Locality	Latitude	Longitude
0	Box Hill	3128.0	Eastern Metropolitan	-37.8151	145.1350
1	Bulleen	3105.0	Eastern Metropolitan	-37.7599	145.0849
2	Doncaster	3108.0	Eastern Metropolitan	-37.7779	145.1270
3	Eaglemont	3084.0	Eastern Metropolitan	-37.7626	145.0570
4	Heidelberg Heights	3081.0	Eastern Metropolitan	-37.7501	145.0505

I also extracted Mean housing prices for the Suburbs of Melbourne and Sydney gather in the dataframes

	Suburb	Mean_house_price_in_K
0	Abbotsford	1126.410256
1	Aberfeldie	1462.965517
2	Airport West	766.807712
3	Albanvale	540.333333
4	Albert Park	1950.985294

Prepare Sydney Data

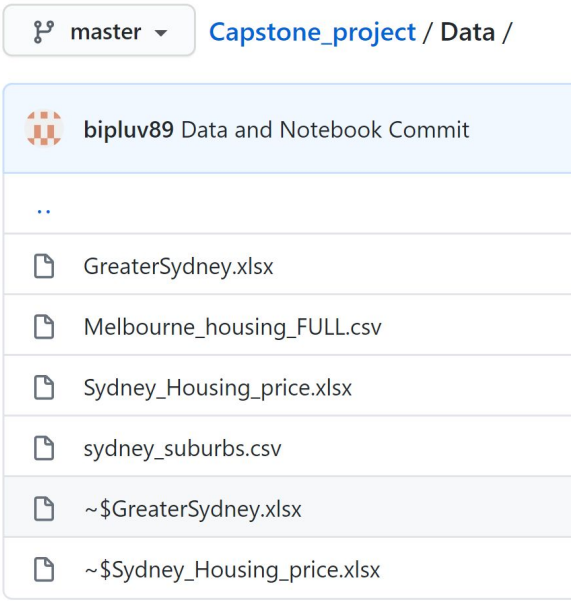
Sydney Suburb data will be prepared from 2 different ways. There is a Kaggle dataset for Sydney suburbs(mentioned in the Data Acquisition). This dataset only has the Suburb name and Coordinates. So, I am using the Australia post API to fetch Suburb's Locality and Postcode data. To limit the number of suburbs from Sydney, I have got a list of Suburb from Greater Sydney Area and I would be using the same to filter the Suburb data for Sydney. Reason for limiting this is same as mentioned for Melbourne

	Suburb	Postcode	Locality	Latitude	Longitude
0	Ashfield	1800	ST GEORGE	-33.889478	151.127412
2	Lane Cove	1595	CHATSWOOD	-33.816067	151.167820
4	Leichhardt	2040	LEICHHARDT	-33.881933	151.155867
5	Marrickville	1475	LEICHHARDT	-33.910385	151.155691
7	Mosman	2088	CHATSWOOD	-33.830135	151.244766

	Suburb	Mean_house_price_in_K
0	Ashfield	844.0
1	Botany Bay	1120.2
2	Lane Cove	885.4
3	Leichhardt	819.0
4	Marrickville	1224.1

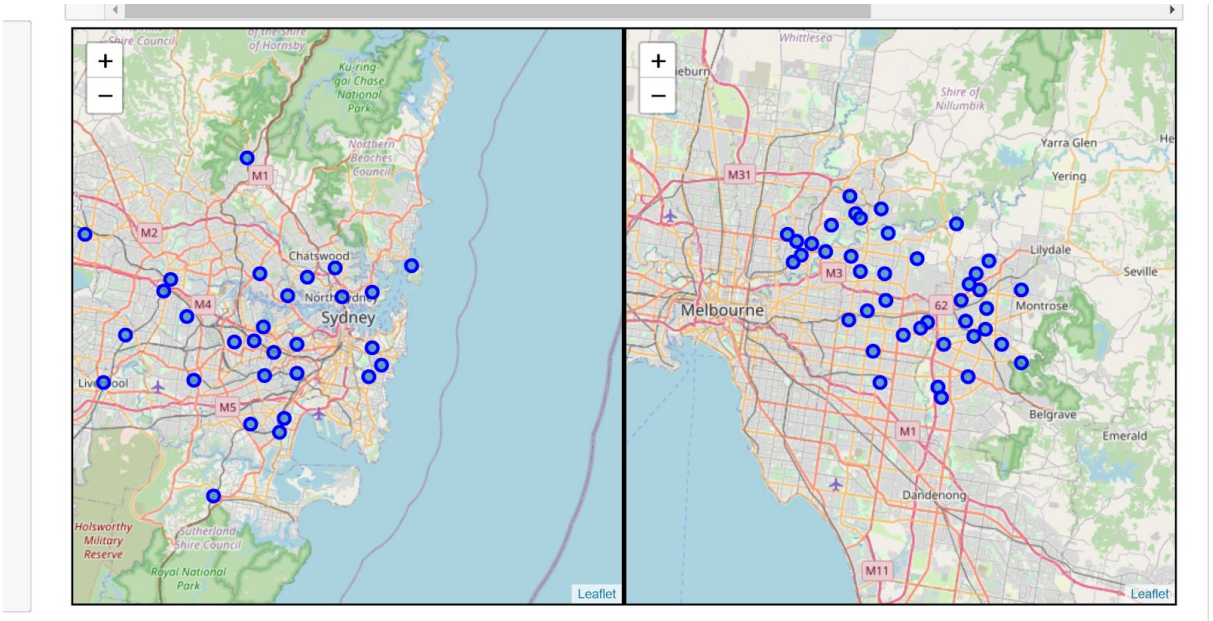
Data Analysis Methodology:

Master data ,i.e. data for Sydney and Melbourne suburb, is stored in Github Repository and data wrangling logic is run by extracting and saving data into GitHub.



Visualization on Map:

Python's Folium Library is used to visualize the suburbs in the Map.



I will be using Foursquare's Venues API to do analysis of finding the top 10 most common venues in each suburb.
The logic Foursquare is to find out which venues or places of interest are most common in each suburb. All the suburbs are expected to have a number of venues which makes the

user decide how the suburb's lifestyle measures. Out of a number of venues, the analysis to showcase top 10 most common venues will help users understand if user's points of interest match with the most common venues in the suburb.
Here below is a display of the number of venues in each suburb:

Suburb	Suburb Latitude	Suburb Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Ashfield	41	41	41	41	41	41
Auburn	3	3	3	3	3	3
Bankstown	50	50	50	50	50	50
Blacktown	25	25	25	25	25	25
Burwood	42	42	42	42	42	42
Camden	10	10	10	10	10	10
Campbelltown	4	4	4	4	4	4
Canada Bay	7	7	7	7	7	7
Canterbury	18	18	18	18	18	18
Fairfield	14	14	14	14	14	14
Holroyd	5	5	5	5	5	5
Hornsby	9	9	9	9	9	9
Hunters Hill	8	8	8	8	8	8
Hurstville	2	2	2	2	2	2
Kogarah	11	11	11	11	11	11

After analysing the the suburbs the top 10 most common venues will be displayed to users as show below:

	Suburb	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bayswater	Vietnamese Restaurant	Malay Restaurant	Video Store	Grocery Store	Thai Restaurant	Dumpling Restaurant	Supermarket	Cantonese Restaurant	Whisky Bar	Dim Sum Restaurant
1	Bayswater North	Sporting Goods Shop	Convenience Store	Ice Cream Shop	Gas Station	Pizza Place	Football Stadium	Deli / Bodega	Dim Sum Restaurant	Dumpling Restaurant	Electronics Store
2	Box Hill	Hotel	Dance Studio	Gym / Fitness Center	Grocery Store	Cricket Ground	Australian Restaurant	Dim Sum Restaurant	Art Gallery	Gym	Asian Restaurant
3	Briar Hill	Gym / Fitness Center	Deli / Bodega	Restaurant	Electronics Store	Whisky Bar	Football Stadium	Dim Sum Restaurant	Dumpling Restaurant	Fast Food Restaurant	Fish & Chips Shop
4	Bulleen	Art Gallery	Park	Sandwich Place	Car Wash	Café	Whisky Bar	Football Stadium	Dumpling Restaurant	Electronics Store	Fast Food Restaurant

Data refining:

As we are comparing Sydney and Melbourne suburbs , so my data analysis I have merged the Suburb details data set, House pricing dataset and Venues data.
This will help in analysing all the suburbs into 1 view which will be further used as a solution for users. I have added a column called City which will be later used to segregate data to visualize the suburbs on the map. The merged data would look below:

	City	Suburb	Mean_house_price_in_K	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
0	Melbourne	Bayswater	766.033333	Vietnamese Restaurant	Supermarket	Malay Restaurant	Cantonese Restaurant	Grocery Store	Thai Restaurant	Dumpling Restaurant	Video Store	Park
1	Melbourne	Bayswater North	750.500000	Ice Cream Shop	Convenience Store	Pizza Place	Sporting Goods Shop	Gas Station	Women's Store	Performing Arts Venue	Pool	Playground
2	Melbourne	Box Hill	1495.067294	Hotel	Dance Studio	Cricket Ground	Gym / Fitness Center	Grocery Store	Women's Store	Performing Arts Venue	Pool	Playground
3	Melbourne	Briar Hill	808.214286	Deli / Bodega	Gym / Fitness Center	Electronics Store	Restaurant	Light Rail Station	Pet Store	Portuguese Restaurant	Italian Restaurant	Pool
4	Melbourne	Bulleen	1150.754054	Park	Café	Car Wash	Sandwich Place	Art Gallery	Japanese Restaurant	Pub	Portuguese Restaurant	Pool

	City	Suburb	Mean_house_price_in_K	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
34	Sydney	Holroyd	469.6	Men's Store	Intersection	Gym	Park	History Museum	Women's Store	Pizza Place	Pharmacy	Pet Store
35	Sydney	Hornsby	585.0	Paper / Office Supplies Store	Café	Pool	Japanese Restaurant	Thrift / Vintage Store	Platform	Bowling Alley	Pet Store	Park
36	Sydney	Liverpool	550.3	Café	Coffee Shop	Fast Food Restaurant	Shopping Mall	Supermarket	Sandwich Place	Seafood Restaurant	Gym	Department Store
37	Sydney	Penrith	599.6	Café	Fast Food Restaurant	Department Store	Thai Restaurant	Performing Arts Venue	Electronics Store	Burger Joint	Sandwich Place	Pub
39	Sydney	Sutherland	973.0	Café	Pub	Thai Restaurant	Sandwich Place	Park	Gym	Farmers Market	Grocery Store	Ramen Restaurant

I have used the unsupervised learning K-means algorithm clustering methodology to cluster the suburbs into different clusters. This algorithm is applied on the merged data for Sydney and Melbourne so that I get an uniform result on all the suburbs. In this way, users will be able to understand which suburbs are similar in characteristics in either city and make an informed decision.
After applying the Machine Learning algorithm, I have merged the cluster labels to the suburb and the view gives users complete knowledge of the suburbs:

	Suburb	Postcode	Locality	Latitude	Longitude	Cluster Labels	City	Mean_house_price_in_K	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
0	Ashfield	1800.0	ST GEORGE	-33.889478	151.127412	0	Sydney	844.0	Dumpling Restaurant	Electronics Store	Supermarket	Japanese Restaurant
2	Lane Cove	1595.0	CHATSWOOD	-33.816067	151.167820	0	Sydney	885.4	Café	Gym	Pizza Place	Chinese Restaurant
4	Leichhardt	2040.0	LEICHHARDT	-33.881933	151.155867	0	Sydney	819.0	Italian Restaurant	Café	Japanese Restaurant	Pub
5	Marrickville	1475.0	LEICHHARDT	-33.910385	151.155691	1	Sydney	1224.1	Vietnamese Restaurant	Café	Ice Cream Shop	Greek Restaurant
7	Mosman	2088.0	CHATSWOOD	-33.830135	151.244766	1	Sydney	1395.7	Café	Bakery	Pub	Clothing Store

It informs user's about Suburb's locality, mean house sales price , top 10 common venues and clusters to which each suburb belongs.

Results:

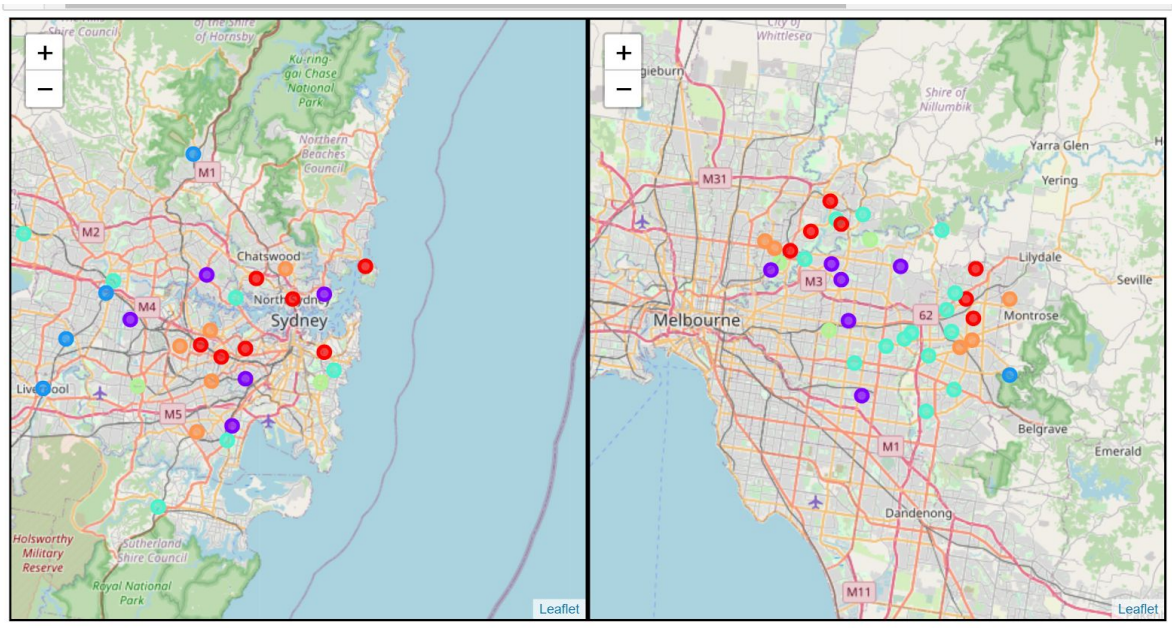
Use of the above analysis: Now, we can create a UI option for users to select the cluster label and view the associated suburbs for each cluster. The same can be visualized on the map as well.

Here below is an example of cluster 5:

	Postcode	Longitude	Cluster Labels	City	Mean_house_price_in_K	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Mos Commo Venu
20	2193.0	151.117015	5	Sydney	703.700000	Park	Platform	Café	Supermarket	Camera Store	Skating Rink	Fast Foo Restaurar
21	2046.0	151.115155	5	Sydney	696.300000	Outdoor Supply Store	Sports Club	Golf Course	Athletics & Sports	Playground	Gym	Coffe Sho
23	2220.0	151.100361	5	Sydney	702.600000	Aquarium	Cricket Ground	Women's Store	Other Nightlife	Pool	Playground	Pizz Plac
34	2135.0	151.081176	5	Sydney	687.200000	Business Service	Women's Store	Pub	Pool	Playground	Pizza Place	Pharmac
35	2068.0	151.200278	5	Sydney	758.200000	Café	Pizza Place	Japanese Restaurant	Bus Station	Malay Restaurant	Gym / Fitness Center	Ga Statio
37	2570.0	150.696170	5	Sydney	671.300000	Supermarket	Café	Park	Bar	Pharmacy	Coffee Shop	Fast Foo Restaurar
4	3081.0	145.050500	5	Melbourne	759.767857	Café	IT Services	Gas Station	Fast Food Restaurant	Pool	Playground	Pizz Plac
5	3081.0	145.039700	5	Melbourne	645.583333	Sculpture Garden	Playground	Pizza Place	Gym	Mobile Phone Shop	Women's Store	Par
14	3153.0	145.260750	5	Melbourne	766.033333	Vietnamese Restaurant	Supermarket	Malay Restaurant	Cantonese Restaurant	Grocery Store	Thai Restaurant	Dumplin Restaurar
15	3153.0	145.274880	5	Melbourne	750.500000	Ice Cream Shop	Convenience Store	Pizza Place	Sporting Goods Shop	Gas Station	Women's Store	Performin Arts Venu

Final visualization:

Analysed Data can be viewed on the map for users to see an overall view of the suburbs and their cluster:



Discussion:

As mentioned in the problem statement, the Data analysis methodology is used to solve the problem of comparing the 2 major cities of Australia. The idea was to find and visualize the major suburbs of Sydney and Melbourne on a map and cluster them into different labels. Users will use the map visualization as a result and filter the data based on cluster labels to understand the analysis.
I have used Python libraries to extract data from different formatted datasets downloaded from kaggle, melbourne and Sydney government websites. The data is prepared and presented in a uniform way so that users can relate the prepared data.
Most important part of the project was the use of the Folium library and Foursquare Venues API. Both of these libraries, helped in the analysis of and visualization of data as an end product for users.
As part of data science machine learning methodology, I have used KMeans clustering algorithm. Kmeans algorithm helped the analysis with a proper segregation of suburbs into different clusters.
Finally, the data was visualized using Folium map library and put together into 2 subplots using the HTML display libraries

Conclusion:

The concept of this analysis is driven for people who are migrating to Australia and need a better understanding of either cities. It gives user's a view of the common venues, housing prices in each suburb and also helps them to relate the suburbs from both the cities in to a common view.
This is just a small use case of this problem statement and the target audience can also be the people living in Australia and wanting to have a better understanding of cities.
The solution can be improvised to give users the opportunity to select different cities to compare and finally filter the cluster labels to view more details of the analysis.
Last but not the least, the methodology is not perfect but I believe the Idea can be extended to come up with a product for not only Australia but also for cities throughout the world.

References:

- Sydney Suburbs Wikipedia and Melbourne Suburb Wikipedia
- kaggle Datasets
- Foursquare Venues API
- Australia Post's Suburb API
- Python Pandas Library
- Folium Library
- Most Importantly - Stackoverflow's help