**Melbourne Suburbs Housing Data Analysis using Foursquare API**

**Introduction**

Melbourne is a popular destination in Australia for immigrants to settle down in and also for locals to move to due to it's vast amount of facilities and ever growing opportunities for career and education. Hence it provides an ideal location for an individual to settle down with a family.

The intention of this project will be to provide a versatile selection scheme with respect to which suburb an individual wants to settle down in. It allows potential buyers to:

1. Select suburbs within the same cluster and have similar facilities in the vicinity, while not experiencing the hustle and bustle of the city.
2. Select suburbs in the same cluster, with a lower housing price with similar facilities in the vicinity
3. Protect themselves against real estate agents scamming them with higher prices by knowing the average price of a given suburb.

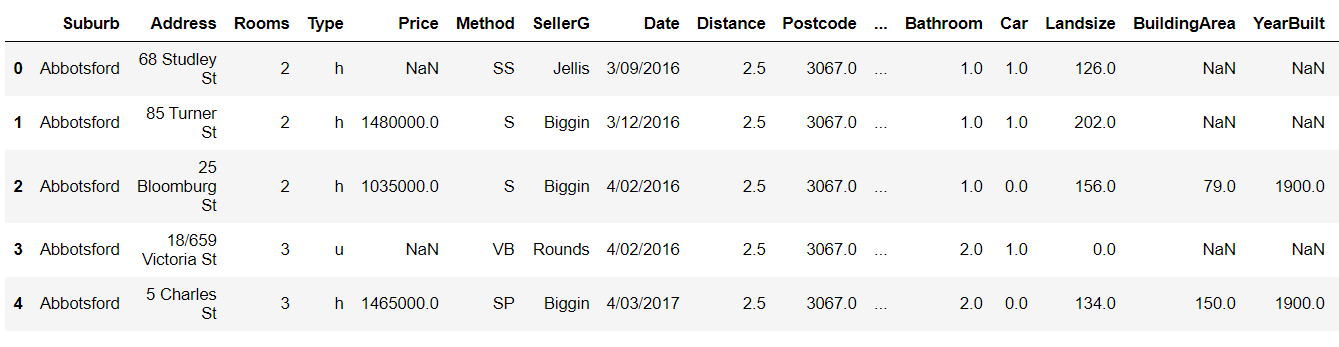
**Data Being Used in this Project**

In this project we aim to use the FourSquare API along with the dataset for Melbourne housing prices taken from this github site (https://raw.githubusercontent.com/nagoya-foundation/r-functions-performance/master/data/Melbourne\_housing\_FULL.csv). The Foursquare API will be used to find the most popular venue categories in each suburb in the dataset. The dataset of Melbourne housing prices will be used to find the average housing price for each suburb.

To plot a clear folium plot showing the variation of prices and the presence of different clusters, GeoJson data from this github site (<https://github.com/tonywr71/GeoJson-Data>) was used.

**Methodology**

The dataset containing the Melbourne Housing Prices consisted of missing values as shown in the following screenshot:



Given that our interest is in the Price of a house, we apply the Groupby method along with a lambda expression to calculate the average price of a suburb and use that value to fill in the NaN values in the Price column.

The following code is used to accomplish the above requirement:



Nevertheless, some NaN values will still occur due to certain suburbs occurring only once in the dataset with a NaN for the price. The mean on 1 NaN value will always be NaN. Hence we drop such instances from the analysis.

The average price of a house in a particular suburb is calculated to obtain the following DataFrame presented in Figure 1:

Figure 1: Part of the DataFrame showing the average price of a suburb



To visualize the location of suburbs in Melbourne, the latitude and longitude of a suburb is found using the Geocoder library and is presented below in Figure 2:

Figure 2: Part of the DataFrame showing the latitude and longitude of a suburb

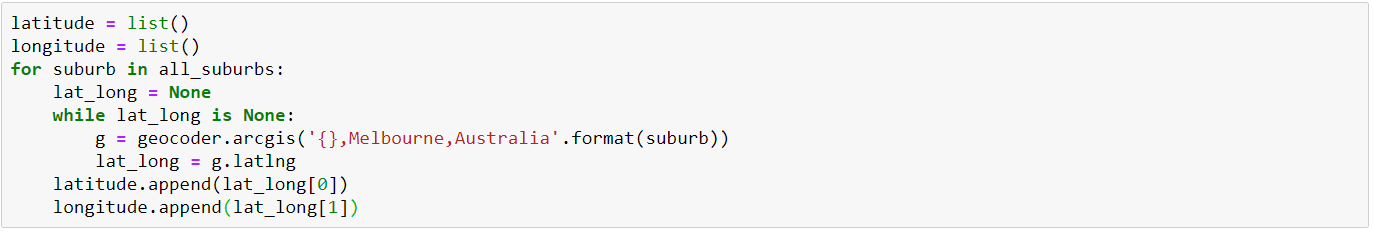
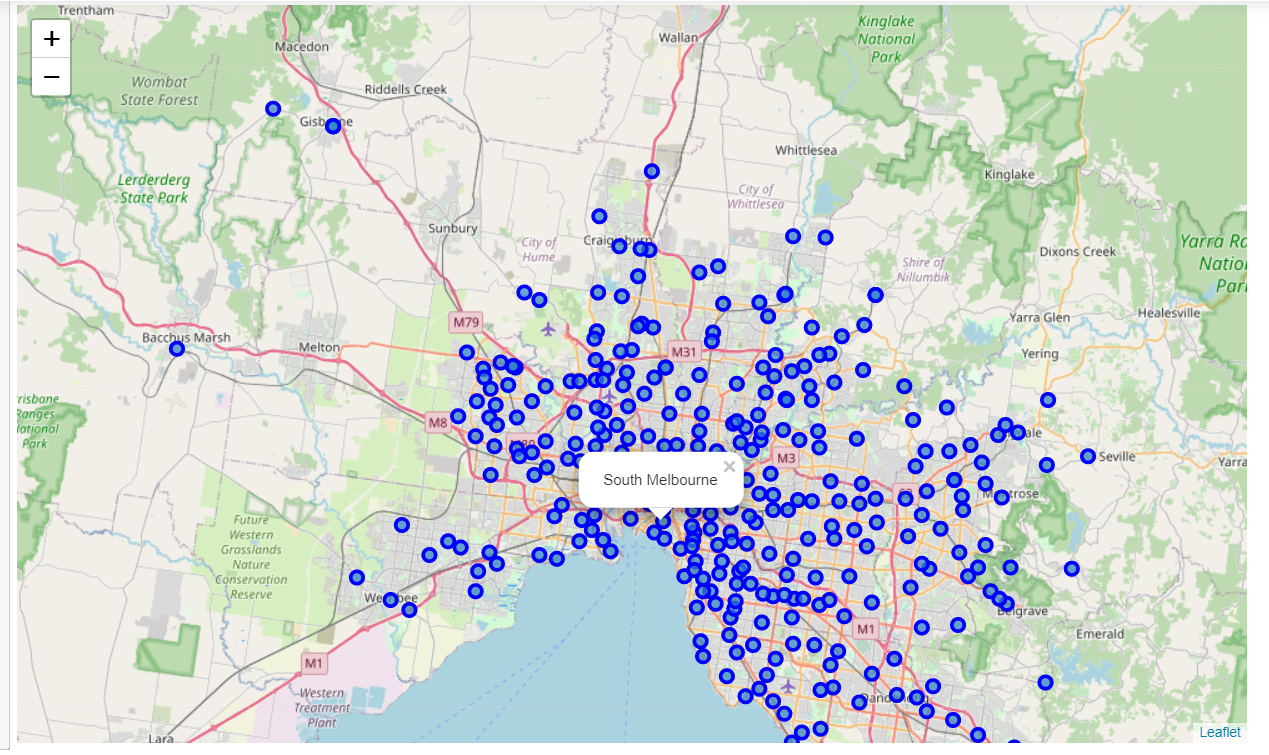
 

Figure 3: Visualizing the suburbs of in melbourne.



**Obtaining the top 50 venues in the vicinity of a suburb**

Using the FourSquare API, requests are made to obtain the json data of all the venues in the vicinity of all the suburbs in the dataset. Figure 4 shows the results of scraping data using the Foursquare API for the Abbortsford suburb.

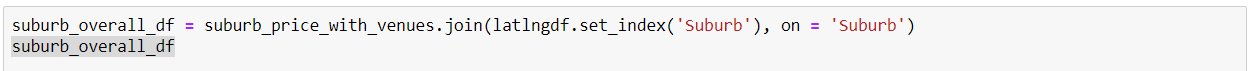
Figure 4: Venues in Abbotsford



One-hot encoding is applied on the Venue categories followed by grouping the data with respect to the Suburb. The Grouped data is averaged across the columns to find the frequency of occurrence of a particular venue category in that suburb.

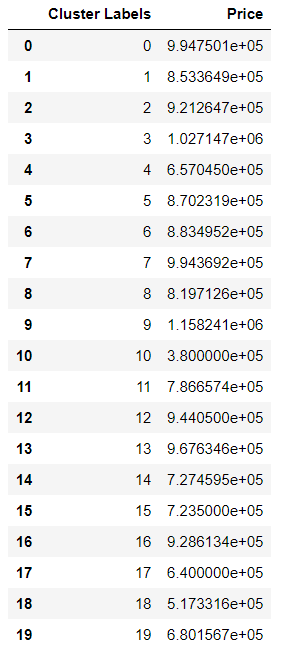
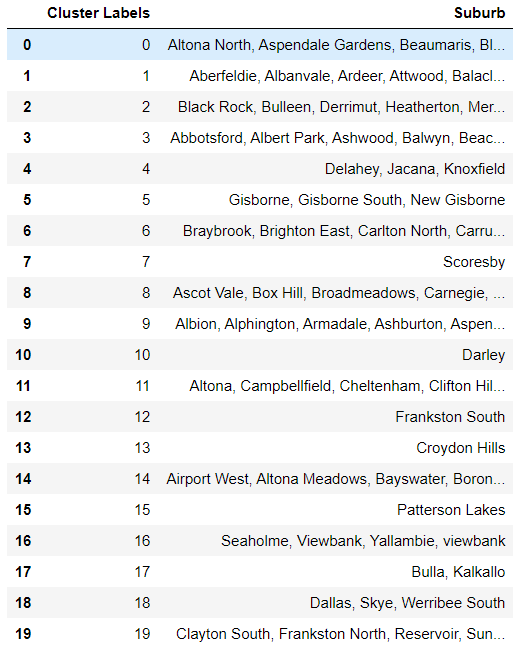
KMeans clustering with 20 clusters is applied to the resulting dataset to obtain the clusters of suburbs which have similar frequency of venue categories occurring in the suburb.

The resulting dataset is merged with the Price of the respective suburb using the following function:

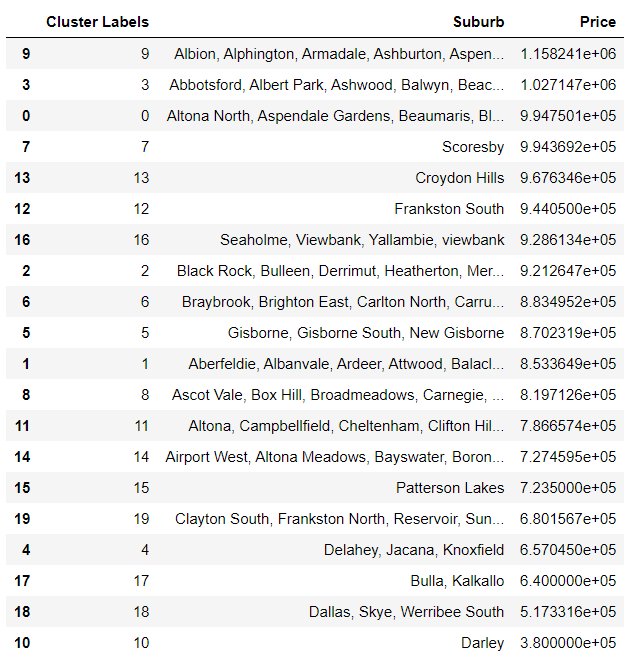


Now we have obtained the required dataset where we have the cluster to which each suburb belongs to and the average price of the suburb. Figures 5,6 and 7 show the average price of each cluster, the suburbs in each cluster and the clusters arranged in descending order showing the cluster with the highest price.

**Figure 6: Average price of a cluster** **Figure 7: Suburbs in a cluster**

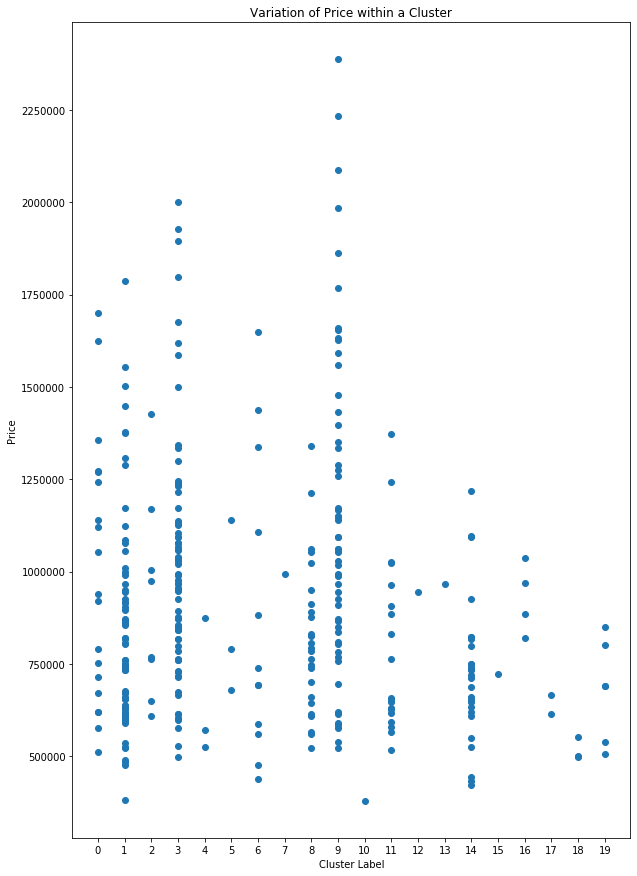
**Figure 8: Clusters arranged in descending order with respect to price**



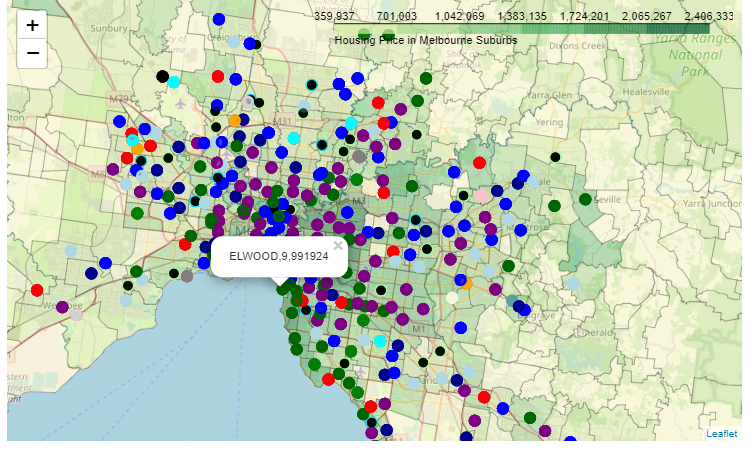
**Results and Discussion**

Figure 9 shows a scatter plot presenting the variation in price within the cluster. Figure 10 shows a folium plot visualizing the price of different suburbs and showing markers of different suburbs while each marker is color coded depending on which cluster they below to. Note that the price of a suburb has been kept independent from the KMeans Clustering. Price was only introduced after clustering was completed.

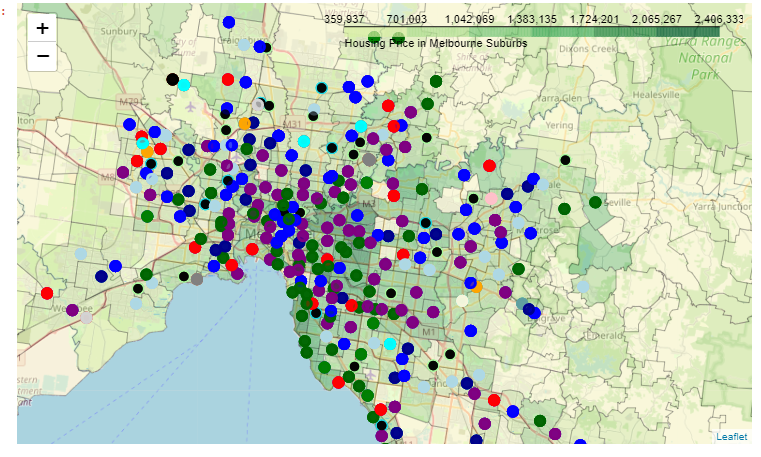
**Figure 9: Variation in price within a cluster**

****

**Figure 10: Folium plot visualizing the price of different suburbs. The pop up label shows suburb name, cluster and average suburb of the suburb**



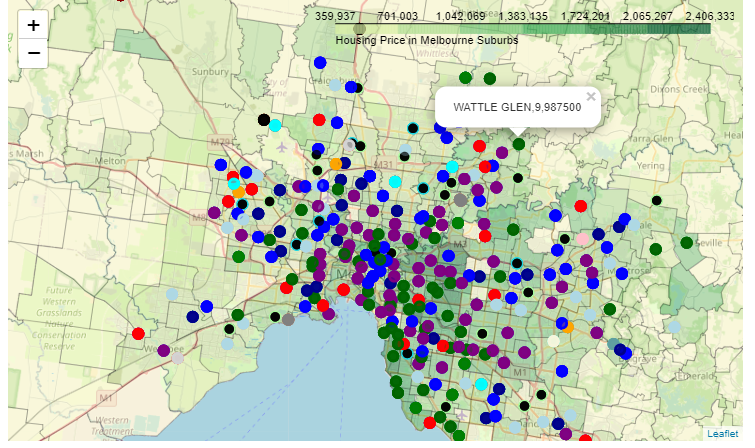
**Figure 11: Folium plot presenting the clusters of suburbs**



As presented in Figure 8, the clusters with highest average housing prices are Clusters 9 and 3. We see that the suburbs closer to central melbourne are predominantly made of these 3 clusters. Hence as expected, closer to the inner-city regions, the price of these suburbs increase. Given that these clusters were processed independently of the price and used the frequency of venue categories, we can conclude that expensive suburbs tend to have similar facilities in their vicinity.

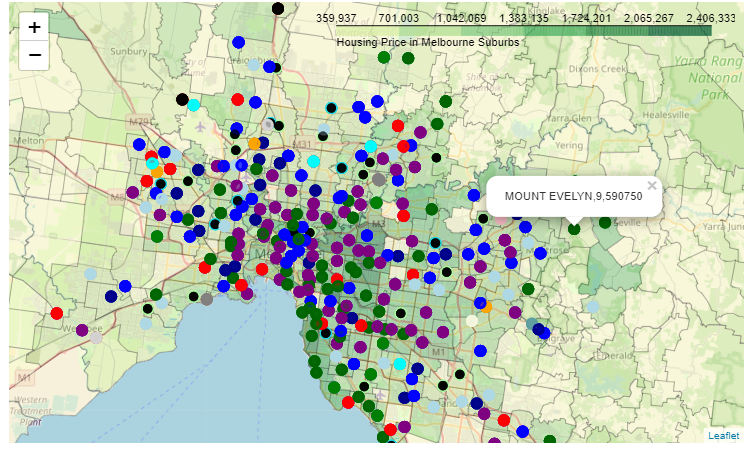
Buyers who want to be away from the inner city, yet have similar facilities may select suburbs within the cluster but away from the inner city as shown in Figure 12.

**Figure 12: A suburb with a similar price but away from the inner city**



Furthermore, as shown in the scatter plot Variation of Price within a Cluster, the price of a house may vary within these clusters itself. This presents an opportunity for buyers as they may opt to go to suburbs with similar facilities in the vicinity, while paying a lower price for a house as shown in Figure 13.

**Figure 13: Lower priced housing within the same cluster**



**Conclusion**

This project has accomplished the goals that were set out at the start of the project which was to:

1. Find clusters of suburbs based on the venue categories in their vicinity
2. Combine them with the average price of those suburbs
3. Present home buyers more flexible options where they may either:
4. Buy homes away from the inner city, yet have similar facilities in their vicinity
5. Buy homes in suburbs with similar facilities to that of inner city suburbs while paying a lower price for the house.