The BOTS (battle of the suburbs)

Louis Huynh

May 9, 2020

1. Introduction

1.1 Background

As a family who is looking to move to Sydney, Australia we want to find the right neighborhood to move into. Location location location is of key importance and currently living remotely makes it harder to do on the ground research. So we want to identify neighborhoods of interest on which to refine our search.

1.2 Problem

For our specific case, this is mainly an exploratory data analysis problem where we want to better understand the area and neighbourhoods to help us refine our search area. It may also be considered a clustering and / or recommendation system as an ideal solution may be to identify different clusters or zones of neighbourhoods and provide a list of recommendations as to which to investigate further. Our initial requirements are that the suburb should be not too far from the central business district, perhaps something like within 20-25 kilometers. It should be a two or three bedroom home and additionally, we would want the median property price to not exceed 800,000 dollars.

1.3 Interest

One of the factors determining location is proximity to good schools (or restaurants). The purpose of this project is to identify neighborhoods that may be prime candidates for moving to based on the number of schools or establishments nearby.

This could also be relevant to business developers who are looking to find neighborhoods that are prime candidates for developing high density residential housing.

2. Data

2.1 Wikipedia - list of suburbs

The data that we will use can be data obtained from Wikipedia in terms of identifying the list of suburbs. We can then supplement with FourSquare data to find the number of and different category of establishments near by. The identification of restuarants and schools can be used as a way to score two different addresses or neighbourhoods when comparing the attractiveness of their relative locations. Plenty of web data is available in terms of auction results including dwelling type, number of bedrooms, bathrooms, car spaces. From this we may also be able to identify different distinct clusters of neighborhood.

As our starting data, we can scrape a list of suburbs from Wikipedia. Then this can be married with auction price data to get average sale prices and addresses of recent properties, which can then be used to identify proximity to the city, number of schools and number of restaurants. From this we can provide a recommeded list of suburbs as a short list from which to begin our property search!

link: https://www.domain.com.au/2-125-euston-road-alexandria-nsw-2015-2016142954

Suburbs are geographical regions in Sydney that identify a neighbourhood. They tend to have different characteristics in terms of property supply and demand as well as property prices. We want to be able to reduce this list of 693 down to perhaps 10 or 20 recommended suburbs that meet other criteria as specified in the introduction and business problem.

2.2 Auction Results - list of addresses

We can get auction results from somewhere like Domain. We can obtain a list of addresses and from these addresses find out more about their location (distance to Sydney, number of restaurants and number of schools.

Sample link to Auction Results: https://www.domain.com.au/auction-results/sydney/2020-05-02

From the summary page we get a list of auction results. Each of these results will have an address and further information such as number of beds, bathrooms and car spaces. We can use this information and query four square to determine the number of nearby schools or restaurants.

Sample link to one specific auction property: https://www.domain.com.au/7-15-17-wyatt-avenue-burwood-nsw-2134-2016164763

This is a townhouse that sold last week for $1.19 m and it's address is 7/15-17 Wyatt Avenue Burwood NSW 2134.

In this specific instance that we take as an example we have the following data and this could potentially be scrapped to get it in bulk.

SOLD - $1,190,000 7/15-17 Wyatt Avenue Burwood NSW 2134

3Beds 1Bath 1Parking 234m² Townhouse

2.3 Four Square Data - list of schools and cafe

For each of the addresses of interest we can obtain the number of restaurant and schools nearby. In order to do so, we must first geocode the address to a set of GPS coordinates and then call the FourSquare api and pass it a search query such as the number of schools. The function query_four_square helps to form us a query and return the results for distances within 1000m. And we see the results of the queries for our example address when running a search query on school and cafe.

The methodology will be as follows:

1. Preprocessing I: We scrape / retrieve list of auction results data from websites like domain.com.au (described above)
2. Preprocessing II: drop null values
3. Exploratory Data Analysis I: explore and filter out values that are not within the Sydney region (we can manually compare to the Wikipedia list of suburbs)
4. Classification I: Run K-means clustering to identifying different groups
5. Post Processing I: Filter out groups that are too far from Sydney
6. Post Processing II: For each group center, find the number of schools and cafes near by and look at the top 3 cluters
7. Post Processing III: Look at suburbs in the top 3 clusters and their average price and sort by ascending prices
8. Post Processing IV: Recommend the top 5 or 10 suburbs.

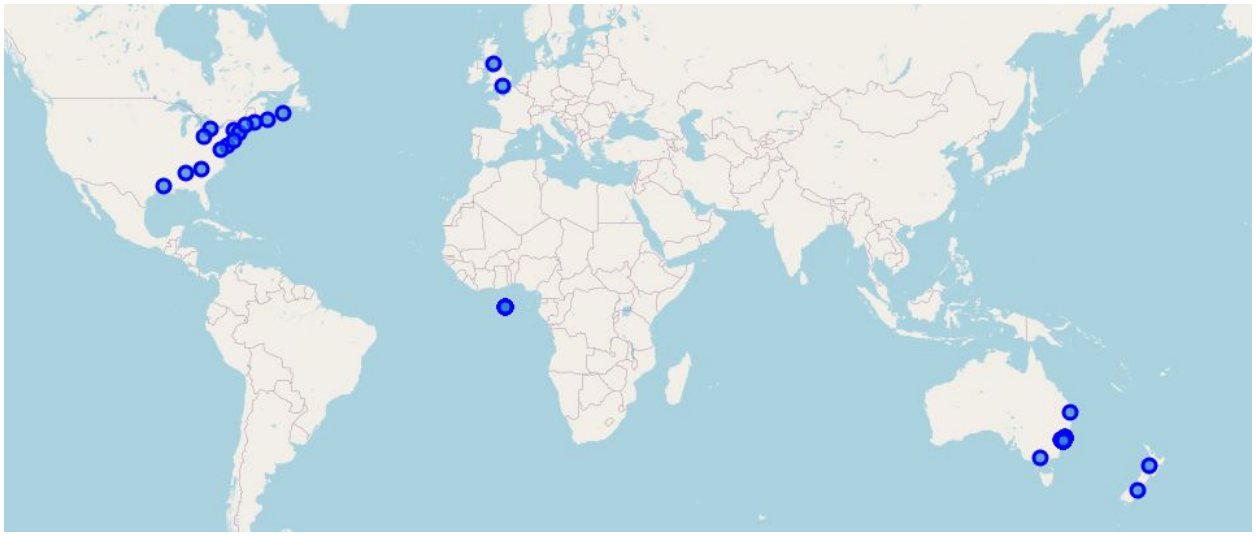
3. Methodology

The overview of the methodology is as follows

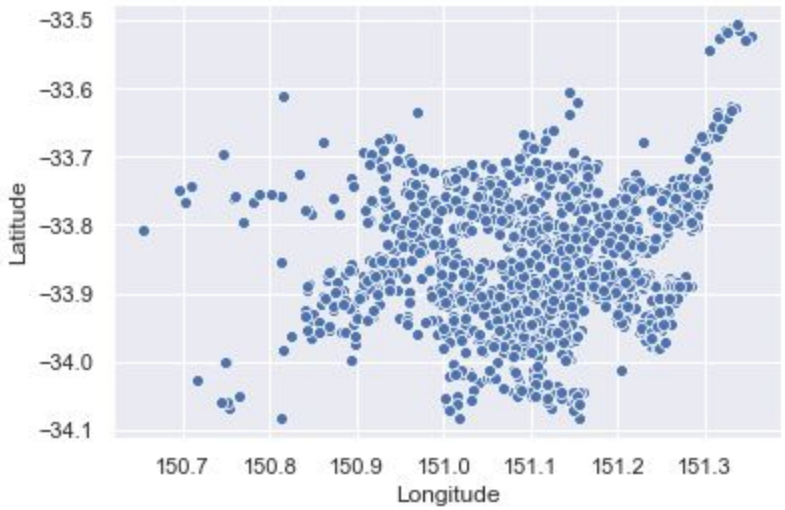
1. Obtain recent auction price results
2. Clean the data by removing entries with bad or missing GPS coordinates
3. Run a model to identify neighbourhood clusters based on GPS coordinates
4. For the center of each neighbourhood
    a. Filter out those that are too far from city center (> 20km)
    b. Find number of nearby schools and cafes
5. Find top 3 neighbourhoods and all the auction records in these neighbourhoods and sort suburbs by average price and recommend the top 10.

3.1 Exploratory Data Analysis

We retrieve our data which includes 11,000 entries and we ignore entries which do not have valid GPS coordinates (this data was not collected for our entries) which results in 2118 remaining entries. Of these remaining entries we would like to visualize on a map and we can see some points that are outside of our desired region (e.g in east coast of US) so we filter these out by applying a boundaries on our GPS coordinates of Latitude between -35 and -33.5 and Longitude greater than 150.6.

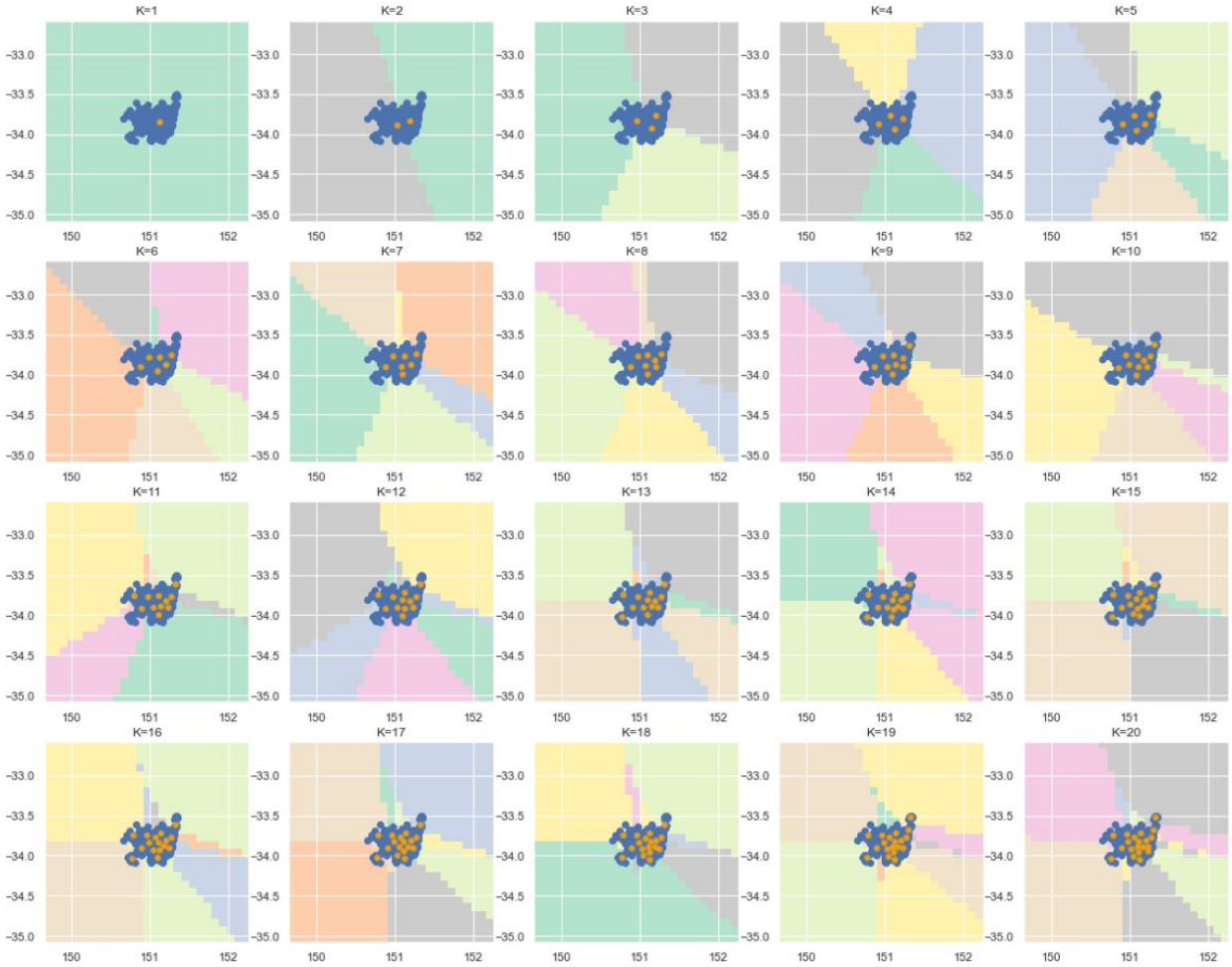


We are then left with 1861 entries which we can further plot and it shows the following:

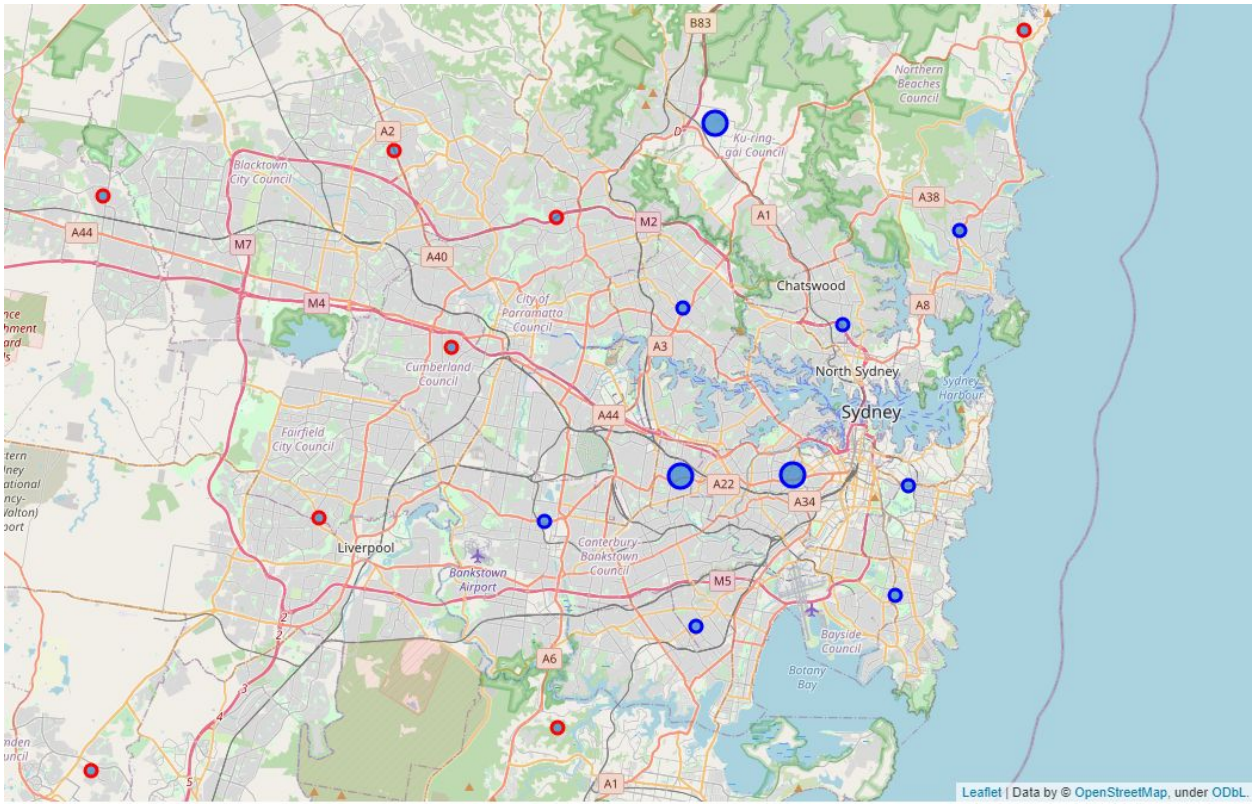


3.2 Clustering

In this phase we then aim to identify different clusters and attempt to find one that is suitable for our exercise. The following plot shows different values of k for clustering:



With 20 clusters then we get the following centroids or centers. We filter out these centroids based on their distance to city. Centers greater than 20km from the city are coloured red and filtered out. Of the remaining clusters, we look up the number of schools and cafes for each centroid and we pick the top three centroids, which is highlighted by the larger circles.



3.3 Ranking

The centroids are identified as follows:

Class	CentroidLatitude	CentroidLongitude	CentroidWithin20km	CentroidDistance	CentroidSchools	CentroidCafes
2	-33.885318	151.170859	True	3.624692	19	47
10	-33.717109	151.126612	True	19.559195	13	5
6	-33.885654	151.106509	True	9.542437	12	41
0	-33.813660	151.200120	True	7.335934	9	50
18	-33.890097	151.238077	True	2.910386	9	50
11	-33.943038	151.230105	True	7.320358	7	17
12	-33.768298	151.267171	True	13.428934	6	27
19	-33.907652	151.028462	True	17.026894	4	17
3	-33.957665	151.115297	True	12.294884	4	3
15	-33.805360	151.108086	True	12.459454	4	15
1	-33.824175	150.974568	False	22.575785	4	1
16	-34.046337	151.123774	False	20.142188	4	8
8	-33.672352	151.304306	False	24.585360	4	13
4	-33.905835	150.898571	False	28.899876	4	2
7	-33.729953	150.942144	False	29.780552	4	13
13	-34.006160	151.035859	False	21.335298	3	0
9	-33.751985	150.774323	False	42.684600	2	1
5	-33.762142	151.035150	False	20.716510	2	5
17	-33.520345	151.331002	False	41.384690	1	12
14	-34.026644	150.767228	False	44.015376	0	0

We go back to our individual auction listing results and look at all suburbs in the top three centroids and then sort these suburbs by their average prices and recommend the top ten.

4. Results

4.1 Top Suburbs

The top ten suburbs end up being:

Suburb	PostCode	KPrice	Latitude	Longitude	Distance to Sydney
St Ives	2075.0	652.222222	-33.726897	151.167501	17.357273
Balmain	2041.0	815.384615	-33.859513	151.180362	3.490000
Stanmore	2048.0	902.909091	-33.891271	151.167791	4.080909
Ashfield	2131.0	1012.882353	-33.886931	151.127924	7.595556
Erskineville	2043.0	1051.363636	-33.903502	151.184830	3.516364
Marrickville	2204.0	1192.122500	-33.911405	151.151615	6.427143
Camperdown	2050.0	1241.339286	-33.888191	151.175908	3.266429
Newtown	2042.0	1410.125000	-33.900790	151.176798	3.862941
Strathfield	2135.0	2053.833333	-33.879597	151.082644	11.753810
Turramurra	2074.0	2434.428571	-33.732508	151.128005	17.983636

There are suburbs which are not too far from the Sydney CBD nor where existing family live and it boasts a high number of schools and cafes. This helps to provide a way to narrow down suburbs systematically and helps to prevent the overlooking of potential good candidate subjects. For example, we see that the median price for a home is St Ives is approximately 650k vs in Strathfield which is just over 2 million. There will indeed be other explanatory features which explain such differences (e.g number of bedrooms, internal area, land size etc) but at least we have a high level starting point on which we can narrow our search.

5. Discussion

5.1 Improvements

A number of parameters were assumed in this case. Some limitations or potential improvements include:

- Looking at a longer dated period of history or auction results

- Doing some analysis on optimal number of K for the k-means clustering (e.g elbow test)
- Tweaking various levels of distance when using the four square api to determine the number of close schools and restaurants
- Exploring the user ratings or scores of schools / restaurants
- Taking into consideration the different dwelling types (e.g house with land, town house vs apartment), internal area size, number of bed and bathrooms etc

5.2 Observations:

- Auction results go through different periods of activity, some periods are a lot more quiet whilst other periods have high demand, this may affect the price
- The four square API does not always seem to return all results and it may be worth exploring different keywords or combining with other data sources
- We limit the use of our Four Square API by only making calls the each centroid as we take it as representative of the neighborhood but we could also tweak the parameters and run for a more specific address to compare different streets within a given neighborhood.

6. Conclusion

In terms of using data to provide recommendations based on distance to CBD, schools and cafes within a certain price point, the following report provides a list of 10 recommended suburbs to go on a short list to narrow down a property purchase search. A number of the recommendations make sense to me and were suburbs that I was already considering. However, a number of other suburbs including the top recommendation of St Ives is something that I have not considered and could be worth exploring and it may represent better value compared to other more popular or high demand suburbs.