The Battle of the Neighbourhoods

Applied Data Science Capstone Project - Mike Mander

Choosing where to live in Singapore

Introduction

Moving to a new place to for a job is hard, having to relocate with a young family creates additional concerns and considerations that have to be factored in. Choosing where to live when you arrive in a new place is one of the most important decisions, and usually you know little to nothing about the city, services, transportation, demographics etc. Making a bad choice at this stage can result in long commutes, resentment about the job and an unhappy social/family life. I will seek to utilise publicly available data and data science skills to attempt to answer the question, Where is the best location for me and my family to live?

My project will assume a hypothetical scenario where a family with young children are moving to Singapore for one parent to start a new job at the IBM office. They have been given a housing allowance of $2000, and a salary with some leeway to spend additional money for the right place, if required.

This is a typical scenario for many expats and so I think that any international company's HR department and new expat employees themselves will be interested in the analytical process and the results. The scenario I am going to resolve for is one of a few potential options, and so I will suggest, during the data analysis, where choices can be made that would suit expats in slightly different situations.

Data

Singapore has a wealth of publicly available government data on their website https://data.gov.sg. I have mined this to get the geoJSON data for their neighbourhoods, or planning areas as they call them. This splits Singapore into 55 areas, and I will use a library called Shapely to find the centroid location coordinates of each area. From those I will utilise the FourSquare API to search for venues that are related to children, kids and babies. In the Developer Categories list I have found many venue types that are related to or would be of interest to parents with young children, including playgrounds, parks, baby shops etc. The full list of IDs will be described in the methodology later.

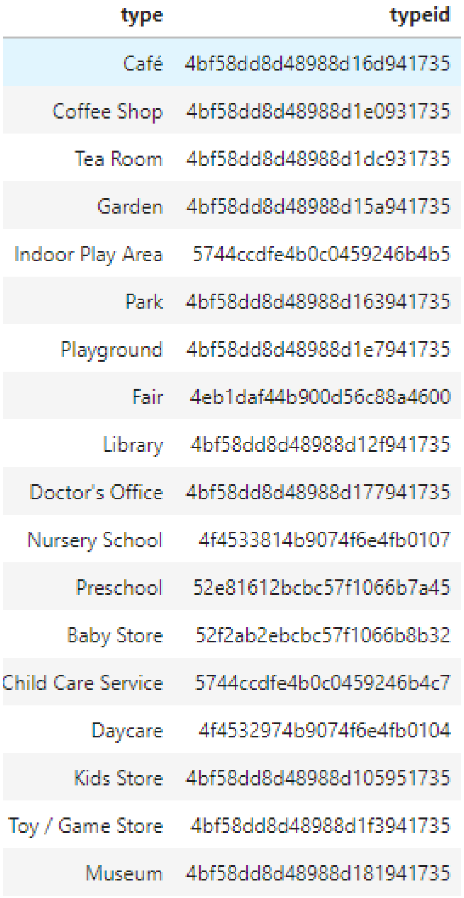
The government website also has a few other datasets that I will be using for the analysis. One data set that I will be using is for an indicative look at the cost of renting, it contains the median rent by area and flat type per quarter from Q2-2005 to Q4 -2018. This data can be interrogated and cleaned to show the prospective expat the expected required budget for the chosen areas as that could potentially influence the decision on where to live.

Another data set I will seek to feature is the ethnic population of each of the planning areas. This is for clustering purposes and I will seek find other areas in Singapore that are similar from a population perspective. This is because expats usually like to live in small clusters, and so a place that’s good for other expats might also be worth considering.

Methodology

Foursquare

Firstly, I interrogated the Foursquare API for all the venues across Singapore that are or could be related to children. I tried to think of places that a parent with children might want to go to or know about locally. Foursquare provide a list of all the categories in their database, and I searched through those noting down the Category and ID’s I required. The list I came up with was:



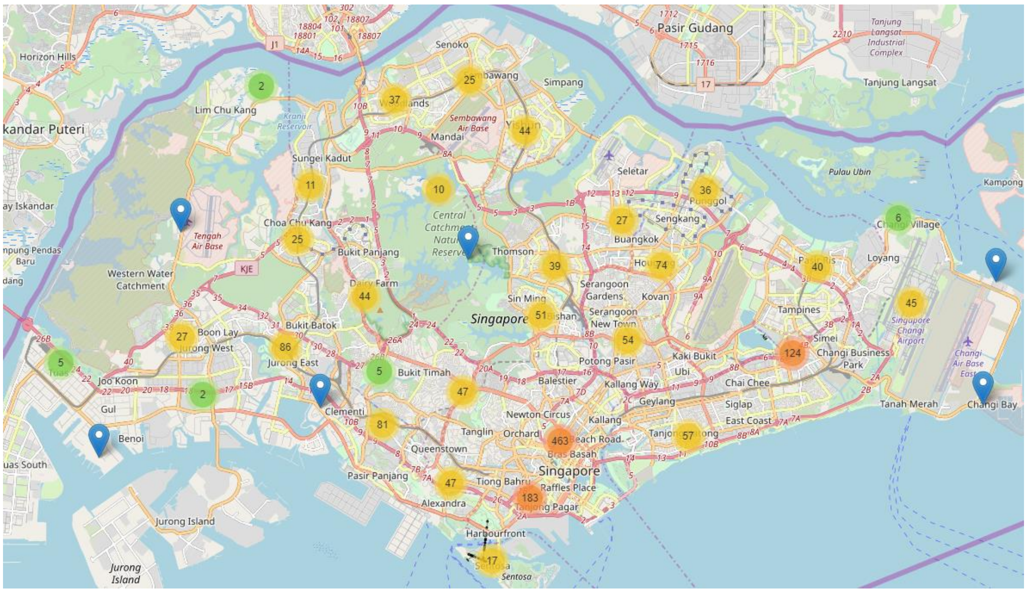
type	typeid
Café	4bf58dd8d48988d16d941735
Coffee Shop	4bf58dd8d48988d1e0931735
Tea Room	4bf58dd8d48988d1dc931735
Garden	4bf58dd8d48988d15a941735
Indoor Play Area	5744ccdfe4b0c0459246b4b5
Park	4bf58dd8d48988d163941735
Playground	4bf58dd8d48988d1e7941735
Fair	4eb1daf44b900d56c88a4600
Library	4bf58dd8d48988d12f941735
Doctor's Office	4bf58dd8d48988d177941735
Nursery School	4f4533814b9074f6e4fb0107
Preschool	52e81612bcbc57f1066b7a45
Baby Store	52f2ab2ebcbc57f1066b8b32
Child Care Service	5744ccdfe4b0c0459246b4c7
Daycare	4f4532974b9074f6e4fb0104
Kids Store	4bf58dd8d48988d105951735
Toy / Game Store	4bf58dd8d48988d1f3941735
Museum	4bf58dd8d48988d181941735

To pull this data out of the API, I needed to configure a few parameters, such as ignoring opening times, and then create a function that processes through the list requesting the venues that match. As there is a limit to the number of venues returns in the response, I needed to write a function that recursively pulls the next batch of venues in automatically. (See Code notebook).

The function then pulls out the name, primary category and location details from the response json, and collates into lists that is converted into a dataframe. At the end we get a fairly large (2041 row) list of all the venues in Singapore that might be relevant to parents.

I want to ensure that the list contains unique venues as there is some scope for the API to pass the same venue a few times dependent on secondary categories. However, as there might well be chain restaurants in the list that would definitely have the same name, I decided to drop all duplicates based on the location coordinates. This reduced the venue list down to 1720 rows.

I then wanted to see where in Singapore this was all located to make sure there was going to be enough of a spread to continue the analysis. I mapped the data as clustered marker on a folium map.

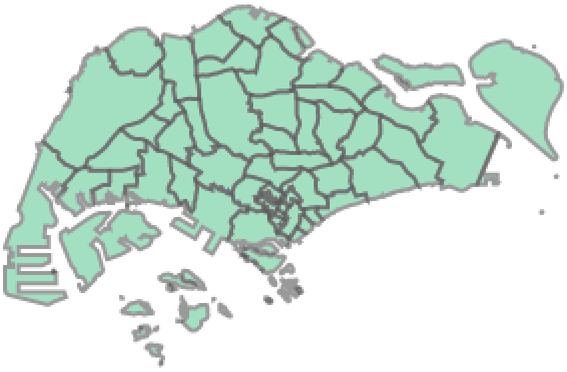


This looked good with a large number in the downtown area as you would expect for shopping and tourism, but still large numbers in other areas of Singapore too.

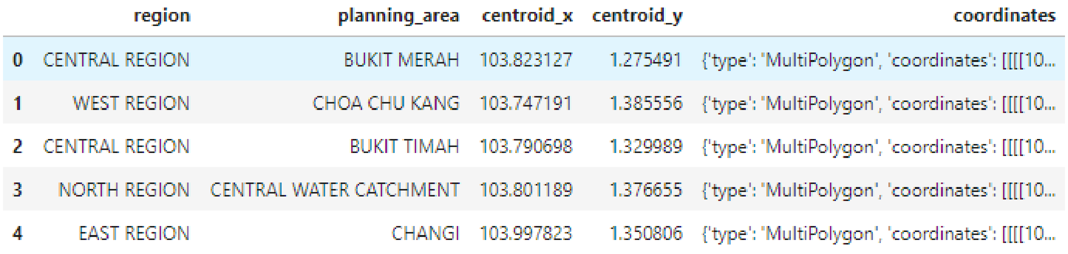
Planning Areas

So now I have all the venues that I need and the coordinate location of them, I need to know which planning area each venue falls under. To do this I am going to use Shapely which is a Python/GIS package. It can take a geojson file and generate polygons with those coordinates, then Points can be checked to see if they fall inside or outside of those polygons.

I downloaded the geojson file of all the planning areas of Singapore and loaded it into Shapely to check it was what I needed.



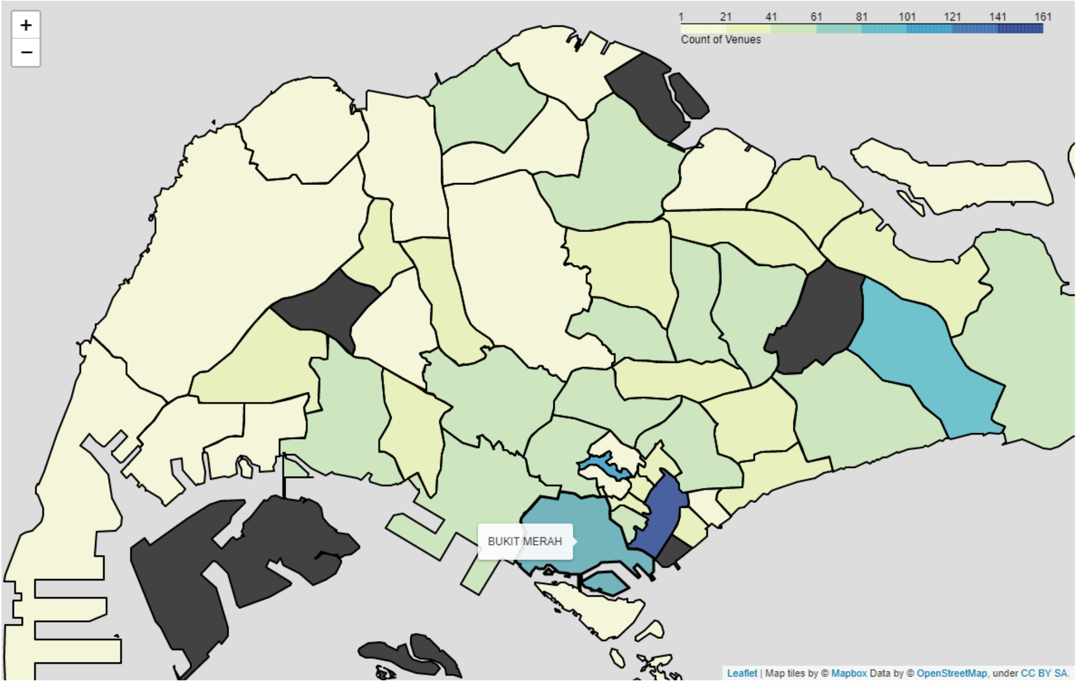
From the raw geojson data I pulled out the Region Name, Planning Area Name and the geometry of the area polygon. I then requested Shapely calculate the centroid location of each area. The resulting dataframe looked like this:



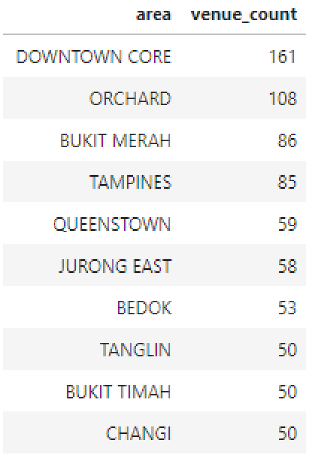
	region	planning_area	centroid_x	centroid_y	coordinates
0	CENTRAL REGION	BUKIT MERAH	103.823127	1.275491	{'type': 'MultiPolygon', 'coordinates': [[[[10...
1	WEST REGION	CHOA CHU KANG	103.747191	1.385556	{'type': 'MultiPolygon', 'coordinates': [[[[10...
2	CENTRAL REGION	BUKIT TIMAH	103.790698	1.329989	{'type': 'MultiPolygon', 'coordinates': [[[[10...
3	NORTH REGION	CENTRAL WATER CATCHMENT	103.801189	1.376655	{'type': 'MultiPolygon', 'coordinates': [[[[10...
4	EAST REGION	CHANGI	103.997823	1.350806	{'type': 'MultiPolygon', 'coordinates': [[[[10...

Now I have a dataframe of the areas and a dataframe of the venues, I need to iterate through all the venues, compare their location against each area polygon and then write back the designated venue area. (See Code notebook)

Now I can produce a venue count for each area, which I will map using a choropleth map to see the areas where the most venues are located.


+
−
1
21
41
61
81
101
121
141
161
Count of Venues
BUKIT MERAH
Leaflet | Map tiles by © Mapbox Data by © OpenStreetMap, under CC BY SA.

This shows that most of the venues are located close to Downtown however there are a few other areas that might also be worth a look. I will get the top 10 areas with highest number of venues.



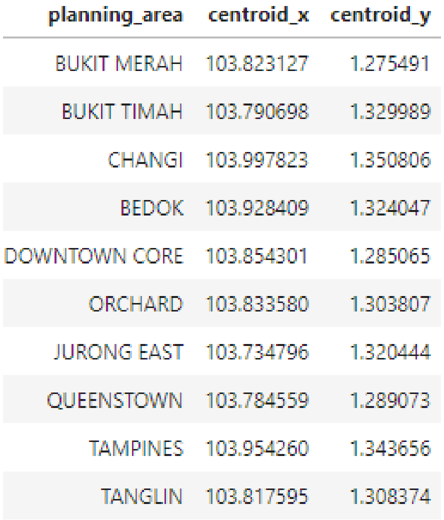
area	venue_count
DOWNTOWN CORE	161
ORCHARD	108
BUKIT MERAH	86
TAMPINES	85
QUEENSTOWN	59
JURONG EAST	58
BEDOK	53
TANGLIN	50
BUKIT TIMAH	50
CHANGI	50

This is a good shortlist to start looking at other factors that might help to influence the decision of where to live.

Commuting Distance

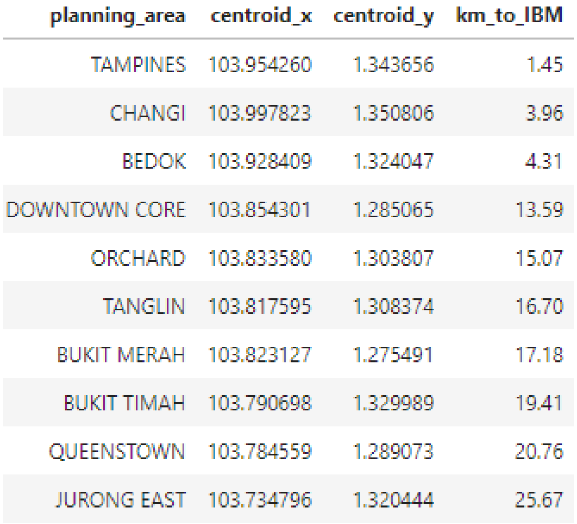
Commuting is a massive drain on your time, it’s difficult to be productive and the longer you spend travelling to and from your place of work the less time you have at home with the family. Sometimes it a price that has to be paid for living in certain areas, however for me I would choose a shorter commute as an important factor in ensuring work/life balance.

I want to look at that list of potential areas and do some distance calculations from each of them to the IBM Office. Firstly, I need to get the location of the IBM Office, so I used Nominatim as a geolocator to give me the coordinates. I then filter the planning area dataframe based on the shortlist to give me the centroid locations of those areas.



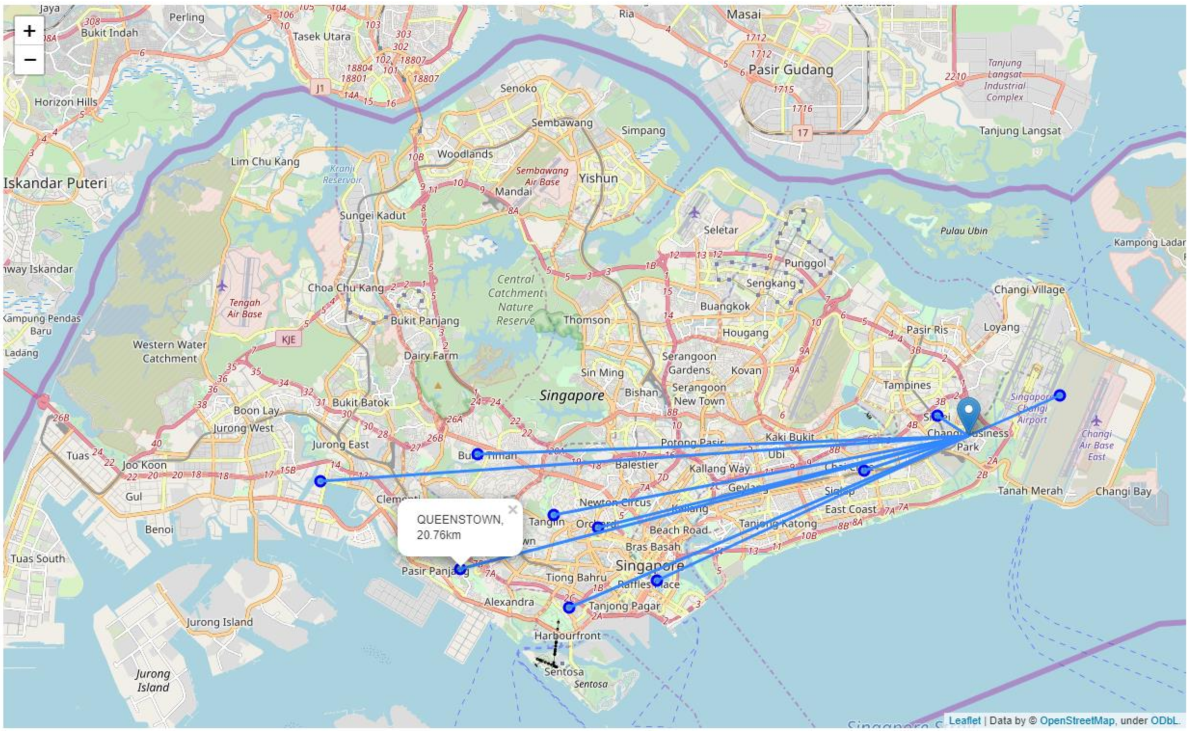
planning_area	centroid_x	centroid_y
BUKIT MERAH	103.823127	1.275491
BUKIT TIMAH	103.790698	1.329989
CHANGI	103.997823	1.350806
BEDOK	103.928409	1.324047
DOWNTOWN CORE	103.854301	1.285065
ORCHARD	103.833580	1.303807
JURONG EAST	103.734796	1.320444
QUEENSTOWN	103.784559	1.289073
TAMPINES	103.954260	1.343656
TANGLIN	103.817595	1.308374

I then load each centroid into a Shapely Point and calculate a distance to the Point set at the IBM Office location. I can then create a new column in the dataframe containing that distance.



planning_area	centroid_x	centroid_y	km_to_IBM
TAMPINES	103.954260	1.343656	1.45
CHANGI	103.997823	1.350806	3.96
BEDOK	103.928409	1.324047	4.31
DOWNTOWN CORE	103.854301	1.285065	13.59
ORCHARD	103.833580	1.303807	15.07
TANGLIN	103.817595	1.308374	16.70
BUKIT MERAH	103.823127	1.275491	17.18
BUKIT TIMAH	103.790698	1.329989	19.41
QUEENSTOWN	103.784559	1.289073	20.76
JURONG EAST	103.734796	1.320444	25.67

Finally for this section, I build a map to see the visual distance of each area centroid to the IBM Office. I know this will only show direct distances rather than commutable distance, but it is still good to see.

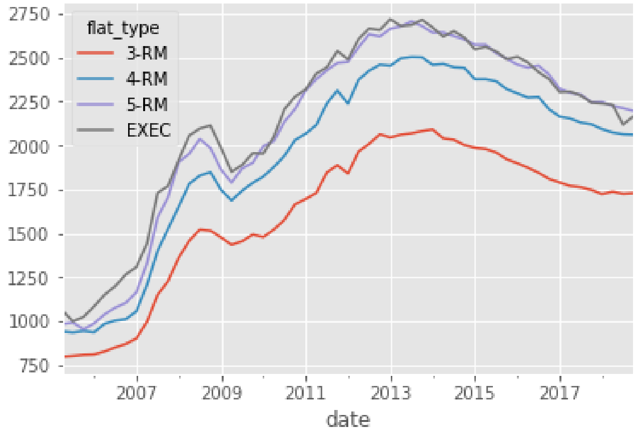


With the office being away from the central downtown area, it obviously would be better living on the East of the island rather than having a long commute from the West.

Rental Prices

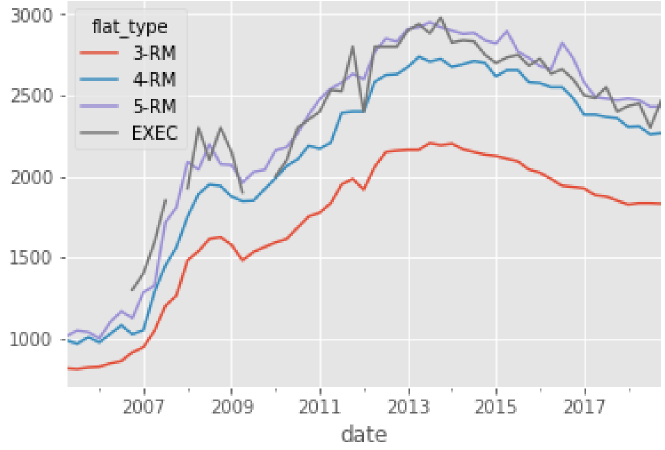
The cost of renting can be a large factor in deciding where to live, I wanted to see whether a nice sized home could be found within the housing allowance budget without having to sacrifice other factors.

The data is quarterly median flat rental costs split up into different sized flats for a number of areas. I plotted the mean price across the areas for each flat type. To see what the Singapore average has been doing.



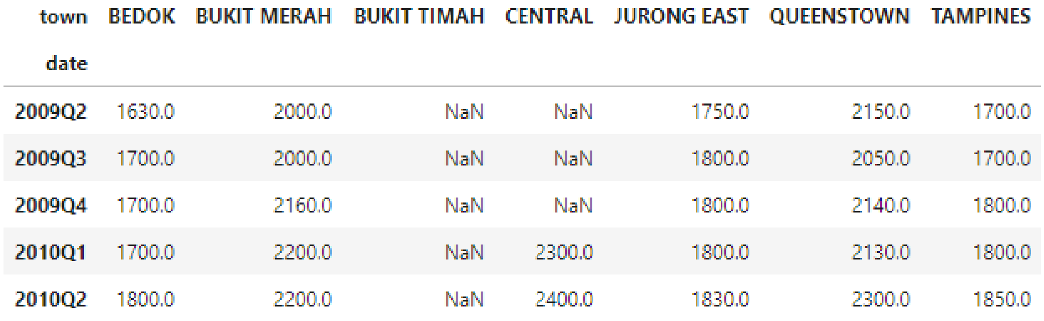
This shows that from 2005 to 2007 there was a large increase in prices, then after the 2008 worldwide financial crash there was a distinct drop but soon after that the prices steadily increased to peak around 2013/2014. Since then there has been a tailing off of the prices and they are now around 20% less than at their peak.

I then filtered the data to only show the areas I had shortlisted, and drew the plot again.



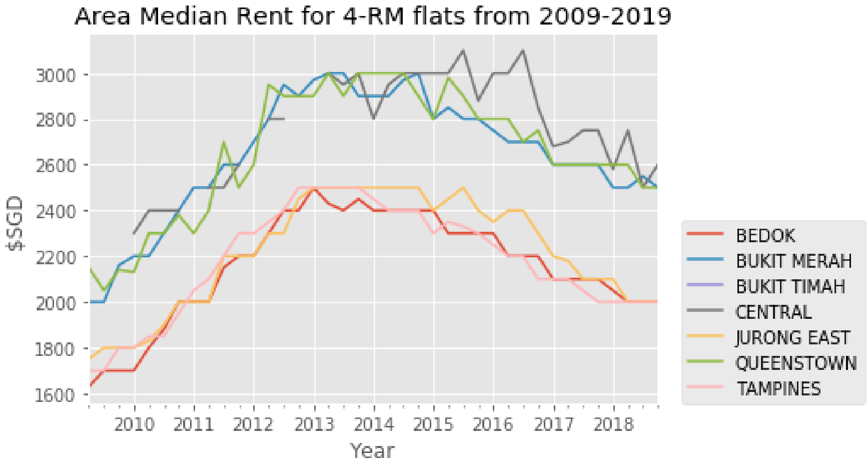
For the areas I am interested in the prices are on average higher than the median for the 4+ room flats.

I wanted to look specifically a 4 room flats as they would offer the space for a young family. So I filtered the dataframe only on them and reduced the time period to 10 years. I then generated a pivot table out of the dataframe so that each area would have its own column of data.



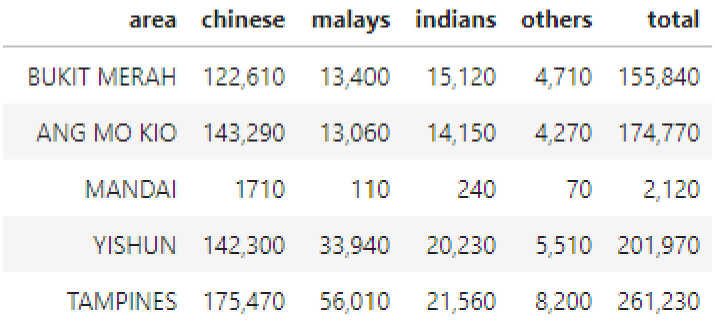
town	BEDOK	BUKIT MERAH	BUKIT TIMAH	CENTRAL	JURONG EAST	QUEENSTOWN	TAMPINES
date							
2009Q2	1630.0	2000.0	NaN	NaN	1750.0	2150.0	1700.0
2009Q3	1700.0	2000.0	NaN	NaN	1800.0	2050.0	1700.0
2009Q4	1700.0	2160.0	NaN	NaN	1800.0	2140.0	1800.0
2010Q1	1700.0	2200.0	NaN	2300.0	1800.0	2130.0	1800.0
2010Q2	1800.0	2200.0	NaN	2400.0	1830.0	2300.0	1850.0

I then plotted this pivot table to see which areas might be better value for money.



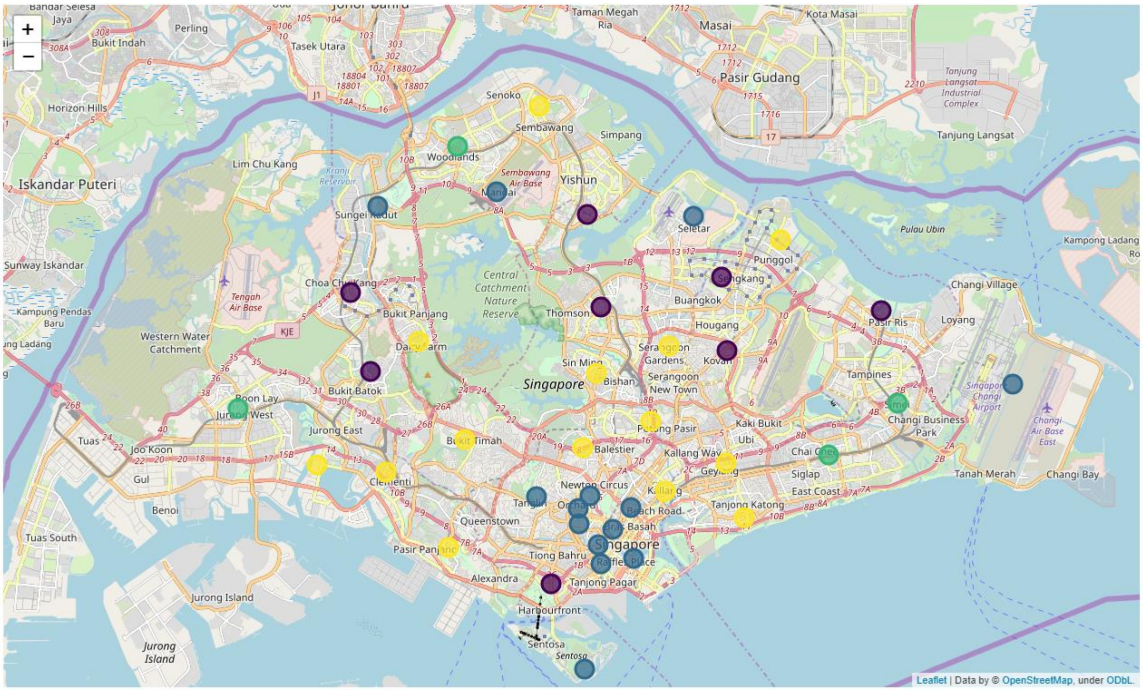
Population clustering

Finally, for the project I wanted to include some data clustering, this was intended to offer some potential alternatives to the obvious best choices, based on other people's housing decisions. The 2015 census allowed a population dataset to be created that listed ethnicity of population in each area. Using the assumption that expat communities often live in fairly close proximity, were there other places with similar numbers of non-local residents that might also be worth a look.



area	chinese	malays	indians	others	total
BUKIT MERAH	122,610	13,400	15,120	4,710	155,840
ANG MO KIO	143,290	13,060	14,150	4,270	174,770
MANDAI	1710	110	240	70	2,120
YISHUN	142,300	33,940	20,230	5,510	201,970
TAMPINES	175,470	56,010	21,560	8,200	261,230

I pulled the dataset into a dataframe and set the index to be the area name. Then after performing some pre-processing I normalised the data values and used K-means clustering to see whether it found similarity. I then added the grouping results to the planning area dataframe and mapped the results.



Results

The results of the analysis show that there are two areas that definitively fit the profile for the scenario. Tampines and Bedok. They are two large areas of population on the south east of Singapore island. They have a total of 138 venues that have been defined as of importance to a young family. They are close to the IBM office, both area centroids are under 5km away. Alongside those factors they also have a median flat rental price within the housing budget. I think either would be suitable candidates for locating the family.

Discussion

The results above have recommended two potential areas for locating a young expat family, I was slightly surprised by the results as I had expected more of the South West areas, like Jurong or Queenstown to be in the running. The locations are probably very well suited for expats, in terms of locality to the airport, being on the MRT line between the airport and the city centre, and also being slightly further out of the city to avoid large crushes of people during rush hours.

As far as the data I used was concerned, I was happy with the Foursquare response and being able to generate the area label from interrogating the geojson file was great. I thought that there could have been more areas for the Rental Prices, I think I may have been lucky in that the areas I was looking at were covered in that dataset, but during my processing I noticed that some area's data was underwhelming and so could have caused problems in different scenarios.

Conclusion

In this study I analysed the publicly available data from Singapore to see if I could use it to determine a good location for a young expat family, moving there to work, to live. I think that it has achieved its aim and would hope that improvements to data and the analysis could allow a model to be built that could weight certain factors to give better results for more varied use cases.