Capstone Project - A Reasonable Location for a New Bookstore in Greater Melbourne (Week2)

by Fendy Gao

date: 2020-Aug-15

Table of Contens

1 Problem Description

Melbourne is the capital and most populous city of the Australian state of Victoria, and the second most populous city in Australia, where I am living. The area of Greater Melbourne area is about 9,993 km2, with a metropolitan area with 31 municipalities. It has a population of about 5 million.

One of my favorite shopping places is bookstore. Due to the impact of on-line sales and electronic publications, it gets more and more difficult to access a regular bookstore now. However, a bookstore is still a very nice place for adults to have some readings and for children to have some learnings. Hence, in this study, I focus myself on bookstores.

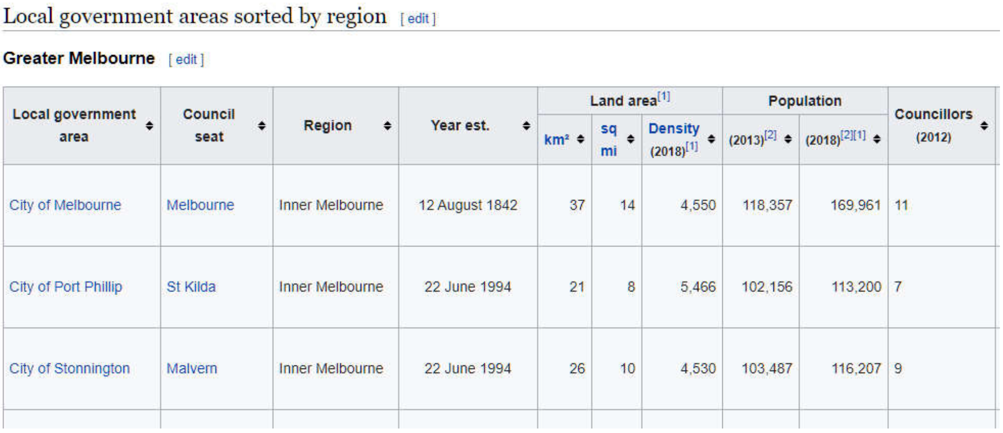
To setup a bookstore would be determined by multiple factors, such as population, transportation, rental rate and etc. Also the existing numbers of bookstores in a certain region would be also imortant. I would like to focus on the factors in the population side. Hence, I'd like to collect data about the number of bookstores in each sub-region of Greater Melbourne Area, the bookstore number vs. population, the accessibility of bookstores to each sub-region, and etc. With all these data analyzed using machine learning technologies, I should be able to yield a recommendation for a location to setup a new bookstore.

In summary, in this project, a study was performed to determine a reasonable location (or locations) to setup a new bookstore (bookstores) in the Greater Melbourne area.

2 Data

2.1 Data of Greater Melbourne

The local government areas (LGAs) of Melbourne are collected from the follow wiki link. I use the LGAs as the basic regions for the coming analysis. The wiki table provided the LGA areas and population in 2018, which is acceptable for the current study.



Local government areas sorted by region [edit]

Greater Melbourne [edit]

Local government area	Council seat	Region	Year est.	Land area[1]			Population		Councillors (2012)
				km²	sq mi	Density (2018)[1]	(2013)[2]	(2018)[2][1]	
City of Melbourne	Melbourne	Inner Melbourne	12 August 1842	37	14	4,550	118,357	169,961	11
City of Port Phillip	St Kilda	Inner Melbourne	22 June 1994	21	8	5,466	102,156	113,200	7
City of Stonnington	Malvern	Inner Melbourne	22 June 1994	26	10	4,530	103,487	116,207	9

Local government areas of Victoria (https://en.wikipedia.org/wiki/Local_government_areas_of_Victoria)

The wiki table of LGAs of Greater Melbourne did not provide any geographical coordinate information. Hence, the following Python package was used to extract geographical coordinate of each LGA.

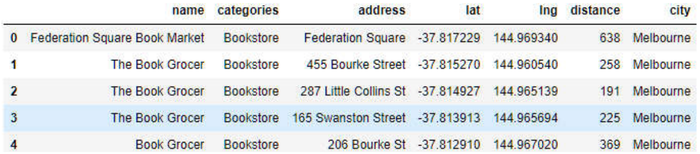
Python Folim (https://github.com/python-visualization/folium)

2.2 Data of Bookstores

In order to find all bookstores in each LGA, Foursquare API was used to find all bookstores using a key word "book" and a reasonable large radius. Some bookstores might be missed from this search, but the current method should be acceptable.

Foursquare API (https://developer.foursquare.com/)

Then, the data was carefully processed. The first action was to remove those venues not in the category of bookstore. The second action was to remove those duplicated venues. The reason was that I used a large radius and certain venues might be found from different LGAs. The bookstore latitudes and longitudes were automatically retrieved via Foursquare API and saved into the dataframe.



	name	categories	address	lat	lng	distance	city
0	Federation Square Book Market	Bookstore	Federation Square	-37.817229	144.969340	638	Melbourne
1	The Book Grocer	Bookstore	455 Bourke Street	-37.815270	144.960540	258	Melbourne
2	The Book Grocer	Bookstore	287 Little Collins St	-37.814927	144.965139	191	Melbourne
3	The Book Grocer	Bookstore	165 Swanston Street	-37.813913	144.965694	225	Melbourne
4	Book Grocer	Bookstore	206 Bourke St	-37.812910	144.967020	369	Melbourne

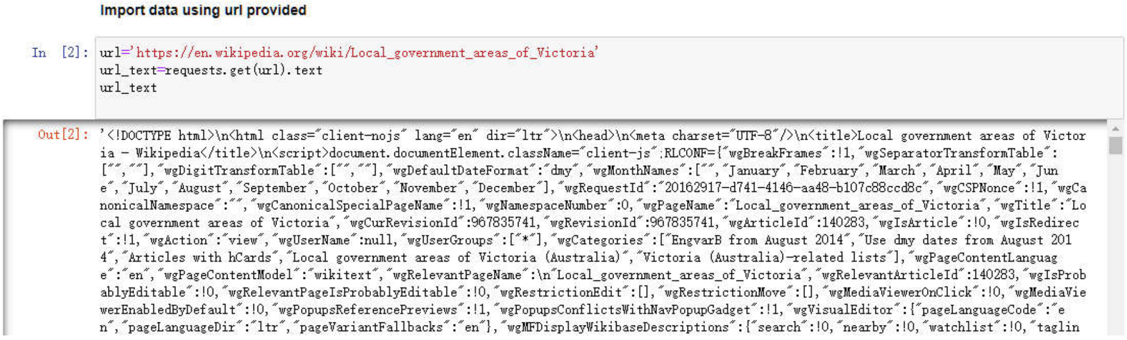
3 Methodology

The following assumption was used to judge where to build a new book store, that a bookstore should be setup where it is needed. The word "needed" is interpretted as that the number of bookstores (or the average number of bookstores per certain population) is less than the average value.

In the study, the number of bookstore in each LGAs were studied, as well as the number of bookstores vs. population, the number of bookstore vs. area, and etc. Then based on the study, the results and recoomendations were presented. Serveral machine learning tools and data plots were used, including bar chart, linear regression, and k-means clustering.

3.1 Process Data of Greater Melbourne

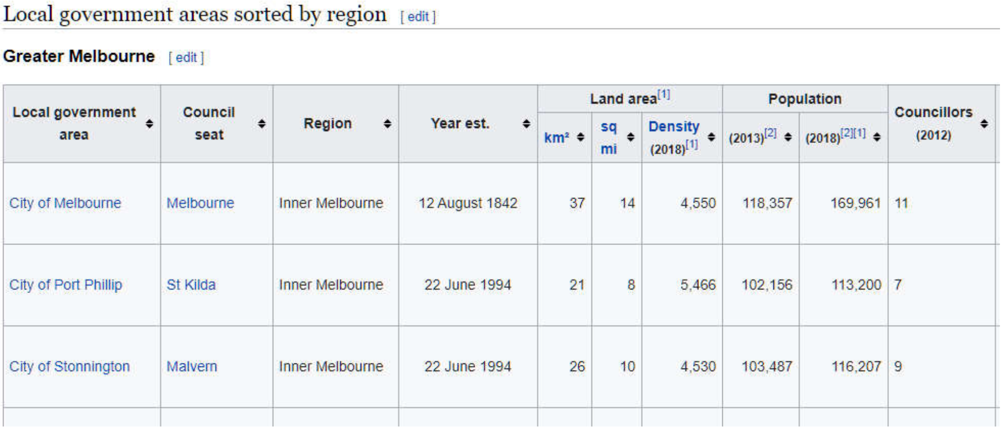
Python *BeautifulSoup* and *requests* module was used to obtain and process Wiki webpage. As shown in the following screeshot, the table includes more data than I needed. During extraction process, all the columns were extracted, but certain columns were not inserted into the final dataframe, such as "Land Area in sq mi" and "Population 2013".



Import data using url provided

```
In  [2]: url='https://en.wikipedia.org/wiki/Local_government_areas_of_Victoria'
         url_text=requests.get(url).text
         url_text
```

```
Out[2]: '<!DOCTYPE html>\n<html class="client-nojs" lang="en" dir="ltr">\n<head>\n<meta charset="UTF-8"/>\n<title>Local government areas of Victor
        ia - Wikipedia</title>\n<script>document.documentElement.className="client-js";RLCONF={"wgBreakFrames":!1,"wgSeparatorTransformTable":
        ["",""],"wgDigitTransformTable":["",""],"wgDefaultDateFormat":"dmy","wgMonthNames":["","January","February","March","April","May","Jun
        e","July","August","September","October","November","December"],"wgRequestId":"20162917-d741-4146-aa48-b107c88ccd8c","wgCSPNonce":!1,"wgCa
        nonicalNamespace":"","wgCanonicalSpecialPageName":!1,"wgNamespaceNumber":0,"wgPageName":"Local_government_areas_of_Victoria","wgTitle":"Lo
        cal government areas of Victoria","wgCurRevisionId":967835741,"wgRevisionId":967835741,"wgArticleId":140283,"wgIsArticle":!0,"wgIsRedirec
        t":!1,"wgAction":"view","wgUserName":null,"wgUserGroups":["*"],"wgCategories":["EngvarB from August 2014","Use dmy dates from August 201
        4","Articles with hCards","Local government areas of Victoria (Australia)","Victoria (Australia)-related lists"],"wgPageContentLanguag
        e":"en","wgPageContentModel":"wikitext","wgRelevantPageName":\n"Local_government_areas_of_Victoria","wgRelevantArticleId":140283,"wgIsProb
        ablyEditable":!0,"wgRelevantPageIsProbablyEditable":!0,"wgRestrictionEdit":[],"wgRestrictionMove":[],"wgMediaViewerOnClick":!0,"wgMediaVie
        werEnabledByDefault":!0,"wgPopupsReferencePreviews":!1,"wgPopupsConflictsWithNavPopupGadget":!1,"wgVisualEditor":{"pageLanguageCode":"e
        n","pageLanguageDir":"ltr","pageVariantFallbacks":"en"},"wgMFDisplayWikibaseDescriptions":{"search":!0,"nearby":!0,"watchlist":!0,"taglin
```

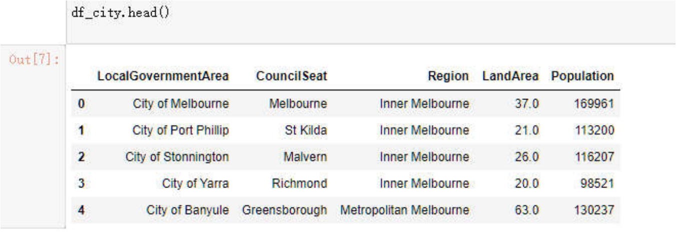


Local government areas sorted by region [edit]

Greater Melbourne [edit]

Local government area	Council seat	Region	Year est.	Land area[1]			Population		Councillors (2012)
				km²	sq mi	Density (2018)[1]	(2013)[2]	(2018)[2][1]	
City of Melbourne	Melbourne	Inner Melbourne	12 August 1842	37	14	4,550	118,357	169,961	11
City of Port Phillip	St Kilda	Inner Melbourne	22 June 1994	21	8	5,466	102,156	113,200	7
City of Stonnington	Malvern	Inner Melbourne	22 June 1994	26	10	4,530	103,487	116,207	9

Since all contents from the table was retrieved as text, the data type of certain columns were converted to the right formats. And the the following dataframe was generated.

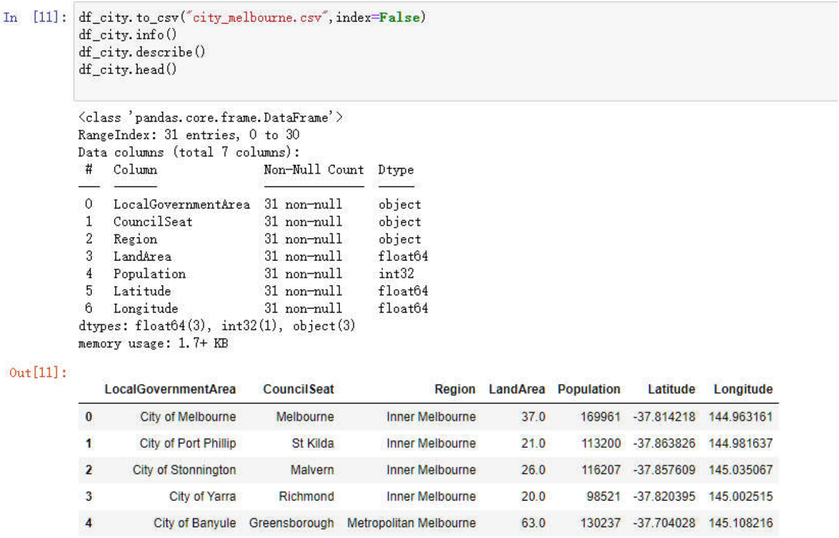


```
df_city.head()
```

Out[7]:

	LocalGovernmentArea	CouncilSeat	Region	LandArea	Population
0	City of Melbourne	Melbourne	Inner Melbourne	37.0	169961
1	City of Port Phillip	St Kilda	Inner Melbourne	21.0	113200
2	City of Stonnington	Malvern	Inner Melbourne	26.0	116207
3	City of Yarra	Richmond	Inner Melbourne	20.0	98521
4	City of Banyule	Greensborough	Metropolitan Melbourne	63.0	130237

As the data in the data frame, the column of CouncilSeat listed all the names of Local Goverment Area (LGA), which were used daily. Hence, in the following step, the names of LGA, combined with the state of Victoria and the name of Australia, were used to obtain the latitudes and longitudes. The Python *geopy* module was used to perform the task. The data were then added into the original dataframe to generate a new dataframe with all the data we needed for the following analysis.



```
In  [11]: df_city.to_csv("city_melbourne.csv",index=False)
          df_city.info()
          df_city.describe()
          df_city.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 31 entries, 0 to 30
Data columns (total 7 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   LocalGovernmentArea  31 non-null     object
 1   CouncilSeat          31 non-null     object
 2   Region               31 non-null     object
 3   LandArea             31 non-null     float64
 4   Population           31 non-null     int32
 5   Latitude             31 non-null     float64
 6   Longitude            31 non-null     float64
dtypes: float64(3), int32(1), object(3)
memory usage: 1.7+ KB
```

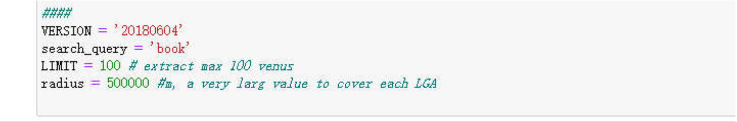
Out[11]:

	LocalGovernmentArea	CouncilSeat	Region	LandArea	Population	Latitude	Longitude
0	City of Melbourne	Melbourne	Inner Melbourne	37.0	169961	-37.814218	144.963161
1	City of Port Phillip	St Kilda	Inner Melbourne	21.0	113200	-37.863826	144.981637
2	City of Stonnington	Malvern	Inner Melbourne	26.0	116207	-37.857609	145.035067
3	City of Yarra	Richmond	Inner Melbourne	20.0	98521	-37.820395	145.002515
4	City of Banyule	Greensborough	Metropolitan Melbourne	63.0	130237	-37.704028	145.108216

3.2 Process of Data of Bookstore

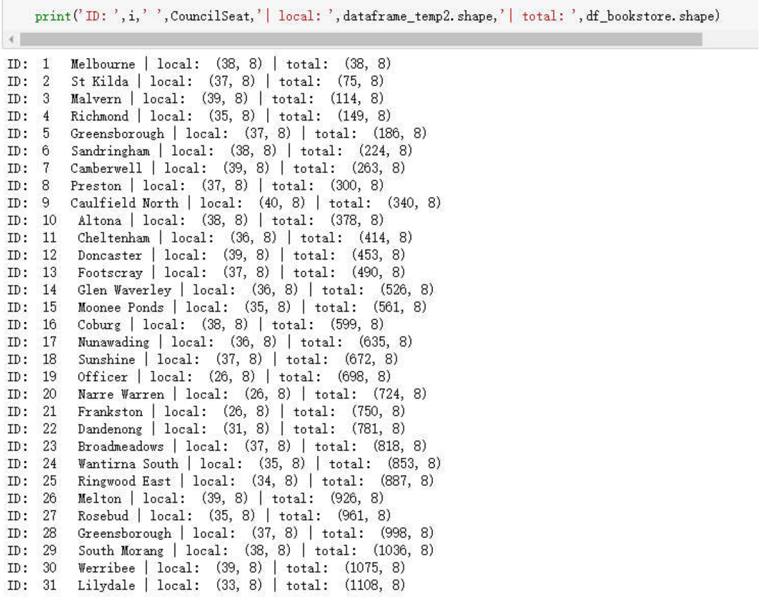
The Foursquare Credentials, Version and other required information were provided based on the query hyperlink with the following format. The version was set to 20180604. The keyword defined in the search was "book". A maximum number of venus retirved was set to 100 with a search radius of 500000, which was pretty large.

```
https://api.foursquare.com/v2/venues/ search ?
client_id= CLIENT_ID &client_secret= CLIENT_SECRET &ll= LATITUDE , LONGITUDE &v= VERSION &query= QUERY &radius= RADIUS &limit= LIMIT
```



```
####
VERSION = '20180604'
search_query = 'book'
LIMIT = 100 # extract max 100 venus
radius = 500000 #m, a very larg value to cover each LGA
```

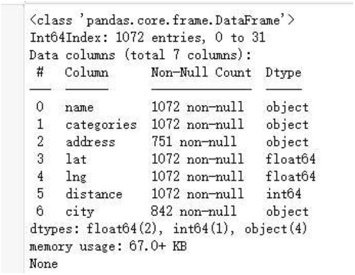
Then, all LGAs of Greater Melbourne were searched and the retrived venues information were appended into a dataframe. The number of venus found in each LGA was less them the value of LIMIT we set above. Hence the value of LIMIT was an acceptable value. Totally, 1109 venus with 'book' in their names were added into the dataframe.



```
print('ID: ',i,' ',CouncilSeat,'| local: ',dataframe_temp2.shape,'| total: ',df_bookstore.shape)
```

```
ID:  1   Melbourne | local:  (38, 8) | total:  (38, 8)
ID:  2   St Kilda | local:  (37, 8) | total:  (75, 8)
ID:  3   Malvern | local:  (39, 8) | total:  (114, 8)
ID:  4   Richmond | local:  (35, 8) | total:  (149, 8)
ID:  5   Greensborough | local:  (37, 8) | total:  (186, 8)
ID:  6   Sandringham | local:  (38, 8) | total:  (224, 8)
ID:  7   Camberwell | local:  (39, 8) | total:  (263, 8)
ID:  8   Preston | local:  (37, 8) | total:  (300, 8)
ID:  9   Caulfield North | local:  (40, 8) | total:  (340, 8)
ID:  10   Altona | local:  (38, 8) | total:  (378, 8)
ID:  11   Cheltenham | local:  (36, 8) | total:  (414, 8)
ID:  12   Doncaster | local:  (39, 8) | total:  (453, 8)
ID:  13   Footscray | local:  (37, 8) | total:  (490, 8)
ID:  14   Glen Waverley | local:  (36, 8) | total:  (526, 8)
ID:  15   Moonee Ponds | local:  (35, 8) | total:  (561, 8)
ID:  16   Coburg | local:  (38, 8) | total:  (599, 8)
ID:  17   Nunawading | local:  (36, 8) | total:  (635, 8)
ID:  18   Sunshine | local:  (37, 8) | total:  (672, 8)
ID:  19   Officer | local:  (26, 8) | total:  (698, 8)
ID:  20   Narre Warren | local:  (26, 8) | total:  (724, 8)
ID:  21   Frankston | local:  (26, 8) | total:  (750, 8)
ID:  22   Dandenong | local:  (31, 8) | total:  (781, 8)
ID:  23   Broadmeadows | local:  (37, 8) | total:  (818, 8)
ID:  24   Wantirna South | local:  (35, 8) | total:  (853, 8)
ID:  25   Ringwood East | local:  (34, 8) | total:  (887, 8)
ID:  26   Melton | local:  (39, 8) | total:  (926, 8)
ID:  27   Rosebud | local:  (35, 8) | total:  (961, 8)
ID:  28   Greensborough | local:  (37, 8) | total:  (998, 8)
ID:  29   South Morang | local:  (38, 8) | total:  (1036, 8)
ID:  30   Werribee | local:  (39, 8) | total:  (1075, 8)
ID:  31   Lilydale | local:  (33, 8) | total:  (1108, 8)
```

The dataframe was then refined by dropping useless column and removing duplicated venues. The bookstore dataframe had the following informaiton.



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1072 entries, 0 to 31
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   name        1072 non-null   object
 1   categories  1072 non-null   object
 2   address     751 non-null    object
 3   lat         1072 non-null   float64
 4   lng         1072 non-null   float64
 5   distance    1072 non-null   int64
 6   city        842 non-null    object
dtypes: float64(2), int64(1), object(4)
memory usage: 67.0+ KB
None
```

Out[6]:

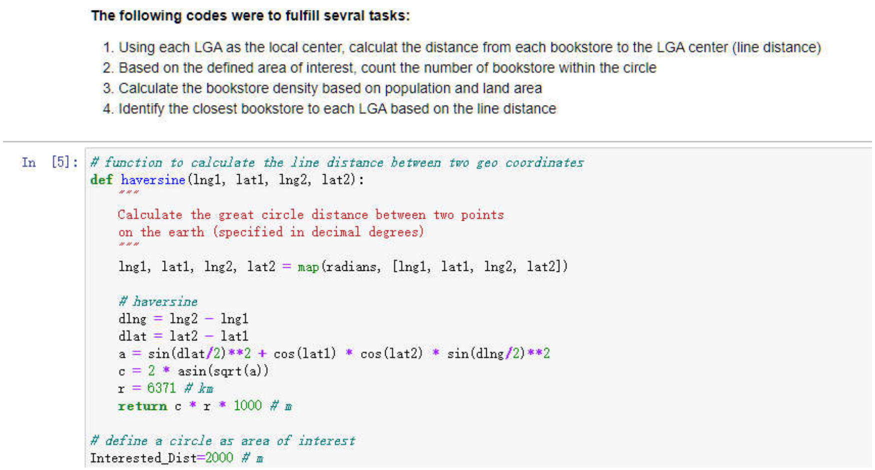
	name	categories	address	lat	lng	distance	city
0	Federation Square Book Market	Bookstore	Federation Square	-37.817229	144.969340	638	Melbourne
1	The Book Grocer	Bookstore	455 Bourke Street	-37.815270	144.960540	258	Melbourne
2	The Book Grocer	Bookstore	287 Little Collins St	-37.814927	144.965139	191	Melbourne
3	The Book Grocer	Bookstore	165 Swanston Street	-37.813913	144.965694	225	Melbourne
4	Book Grocer	Bookstore	206 Bourke St	-37.812910	144.967020	369	Melbourne

3.3 Machine Learning

It is not easy to identify which local region(s) is(are) more suitable for a bookstore. Whether to setup a business in a certain location depends on many factors, such as population, location, tax, resident income level and etc. Hence, a few quantitive indicators were identified and calculated in the study. With these indicators, the possiblity of setting up a bookstore were then further studied.

The first indicator was **scarcity**. Scarcity of bookstore was defined as how many bookstores are shared by the local population or by the local area. The second indicator was **accessibility**, which means how easy a local resident can reach a nearby bookstore. Accesibility was converted into the shortest distance of any bookstore to the local region center.

Based on the defined area of interest and the local region center, the number of bookstores was couted within the circle with a certain radius. In this study, the value of radius was selected as 2000 m, which is about 3-5 minitue driving. The bookstore density was then calculated based on LGA population and land area. The distance from each bookstore to the LGA center (line distance) was calculated. The closest bookstore to each LGA center was then identified.



The following codes were to fulfill sevral tasks:

1. Using each LGA as the local center, calculat the distance from each bookstore to the LGA center (line distance)
2. Based on the defined area of interest, count the number of bookstore within the circle
3. Calculate the bookstore density based on population and land area
4. Identify the closest bookstore to each LGA based on the line distance

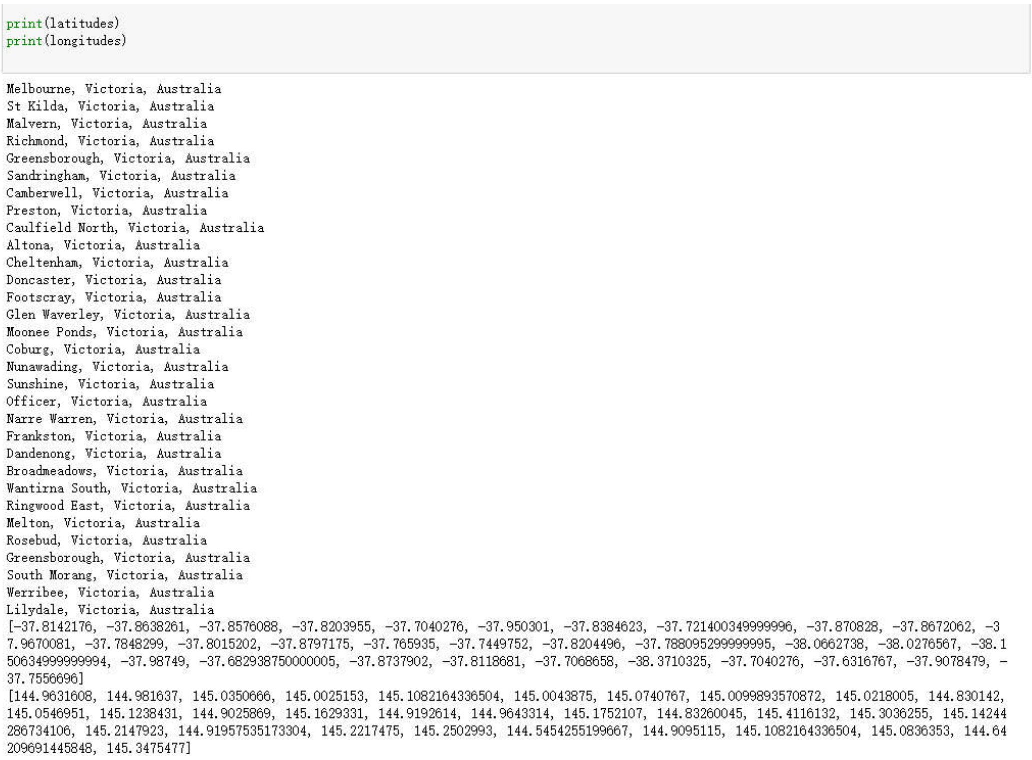
```
In [5]: # function to calculate the line distance between two geo coordinates
        def haversine(lng1, lat1, lng2, lat2):
            """
            Calculate the great circle distance between two points
            on the earth (specified in decimal degrees)
            """
            lng1, lat1, lng2, lat2 = map(radians, [lng1, lat1, lng2, lat2])

            # haversine
            dlng = lng2 - lng1
            dlat = lat2 - lat1
            a = sin(dlat/2)**2 + cos(lat1) * cos(lat2) * sin(dlng/2)**2
            c = 2 * asin(sqrt(a))
            r = 6371 # km
            return c * r * 1000 # m

        # define a circle as area of interest
        Interested_Dist=2000 # m
```

The bookstore density was then plotted based on population and land area of LGA. It was easy to identify the LGA with a low bookstore (even 0) density. The distance of the closest bookstore to each LGA was also plotted. The LGA had a poor **accessibility** to bookstore was also identified.

Scatter plot was use to identify whether there was a correlation between the distance of the closest bookstore with the population or bookstore density. Then, based on the scatter plot and the normalized distance of the closest bookstore and LGA population, k-Means clustering was performed using k value from 2 to 5

4 Results

The local government area of Greater Melbourn were listed and the latitudes and longitudes were also retrieved via python geopy package. The LGA centers were also plot on the map with the radii of circles representing the population in the region.

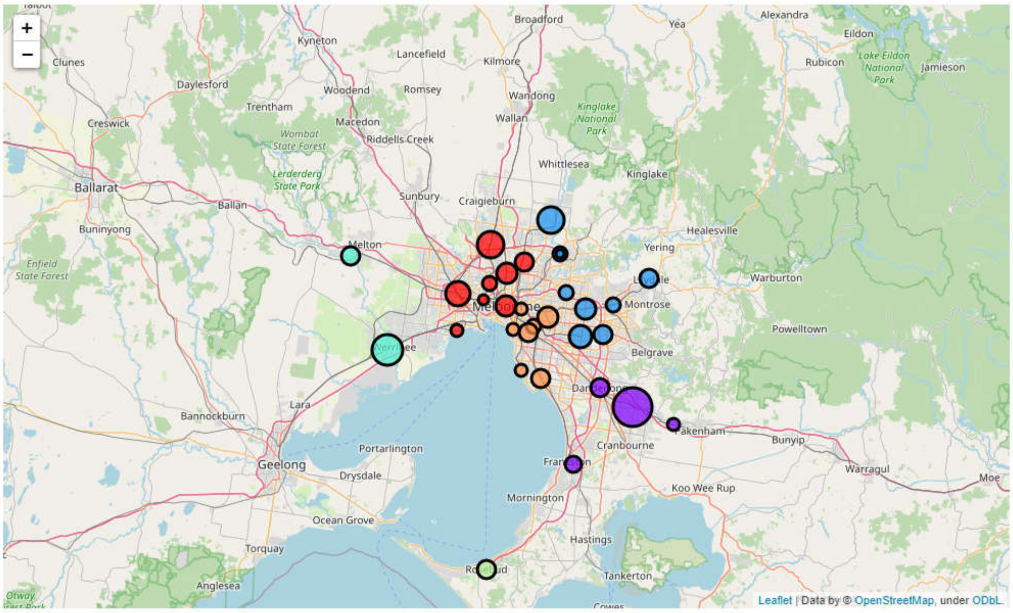


```
print(latitudes)
print(longitudes)
```

```
Melbourne, Victoria, Australia
St Kilda, Victoria, Australia
Malvern, Victoria, Australia
Richmond, Victoria, Australia
Greensborough, Victoria, Australia
Sandringham, Victoria, Australia
Camberwell, Victoria, Australia
Preston, Victoria, Australia
Caulfield North, Victoria, Australia
Altona, Victoria, Australia
Cheltenham, Victoria, Australia
Doncaster, Victoria, Australia
Footscray, Victoria, Australia
Glen Waverley, Victoria, Australia
Moonee Ponds, Victoria, Australia
Coburg, Victoria, Australia
Nunawading, Victoria, Australia
Sunshine, Victoria, Australia
Officer, Victoria, Australia
Narre Warren, Victoria, Australia
Frankston, Victoria, Australia
Dandenong, Victoria, Australia
Broadmeadows, Victoria, Australia
Wantirna South, Victoria, Australia
Ringwood East, Victoria, Australia
Melton, Victoria, Australia
Rosebud, Victoria, Australia
Greensborough, Victoria, Australia
South Morang, Victoria, Australia
Werribee, Victoria, Australia
Lilydale, Victoria, Australia
[-37.8142176, -37.8638261, -37.8576088, -37.8203955, -37.7040276, -37.950301, -37.8384623, -37.721400349999996, -37.870828, -37.8672062, -3
7.9670081, -37.7848299, -37.8015202, -37.8797175, -37.765935, -37.7449752, -37.8204496, -37.788095299999995, -38.0662738, -38.0276567, -38.1
50634999999994, -37.98749, -37.682938750000005, -37.8737902, -37.8118681, -37.7068658, -38.3710325, -37.7040276, -37.6316767, -37.9078479, -
37.7556696]
[144.9631608, 144.981637, 145.0350666, 145.0025153, 145.1082164336504, 145.0043875, 145.0740767, 145.0099893570872, 145.0218005, 144.830142,
145.0546951, 145.1238431, 144.9025869, 145.1629331, 144.9192614, 144.9643314, 145.1752107, 144.83260045, 145.4116132, 145.3036255, 145.14244
286734106, 145.2147923, 144.91957535173304, 145.2217475, 145.2502993, 144.5454255199667, 144.9095115, 145.1082164336504, 145.0836353, 144.64
209691445848, 145.3475477]
```

```
map_clusters
```

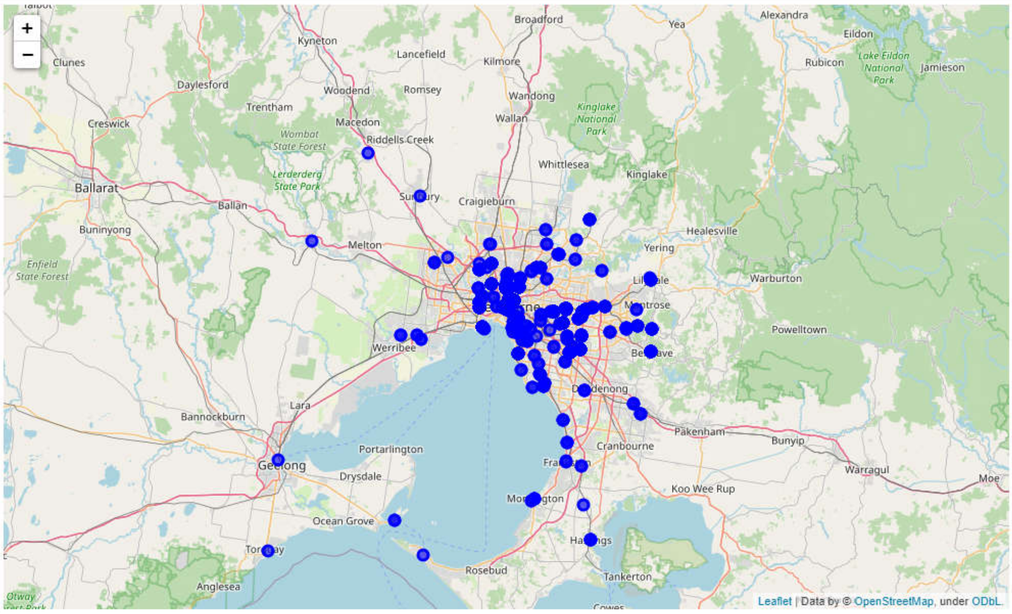
Out[13]:



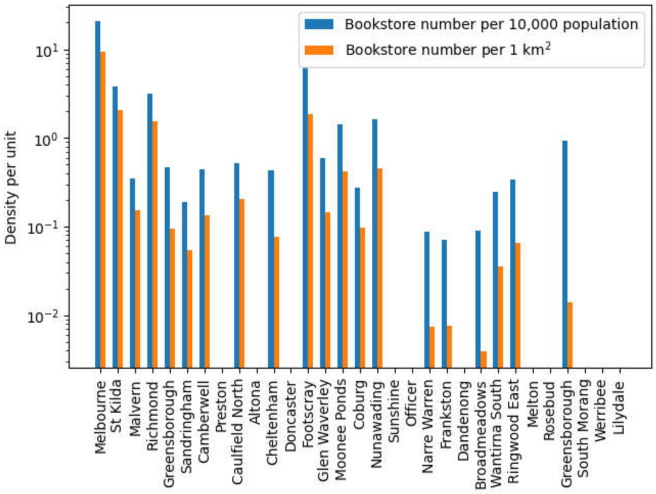
All bookstores searched via Foursqure were also plotted on the map. A few bookstores not in the Greater Melbourne were also found and added in to the dataframe. Those bookstores did not affected the following study, so they were not further removed out from the dataframe.

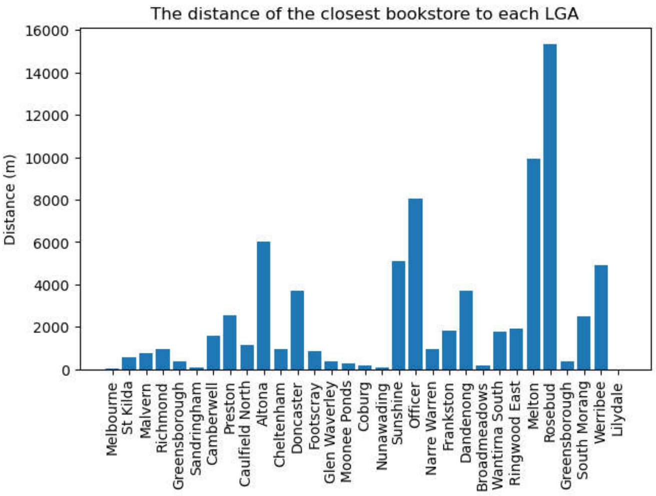
```
venues_map
```

Out[7]:

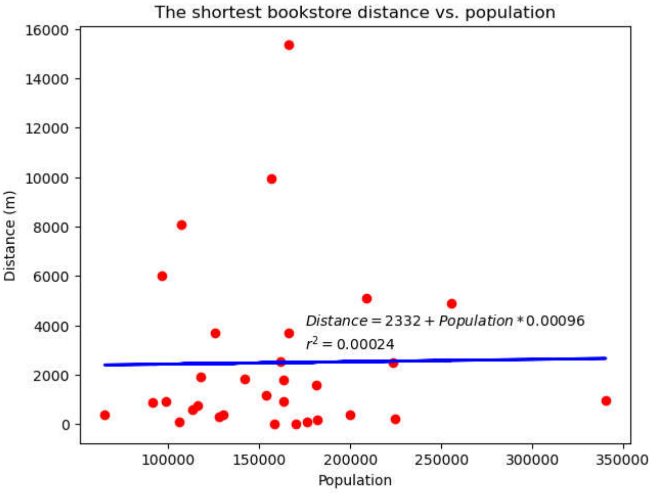


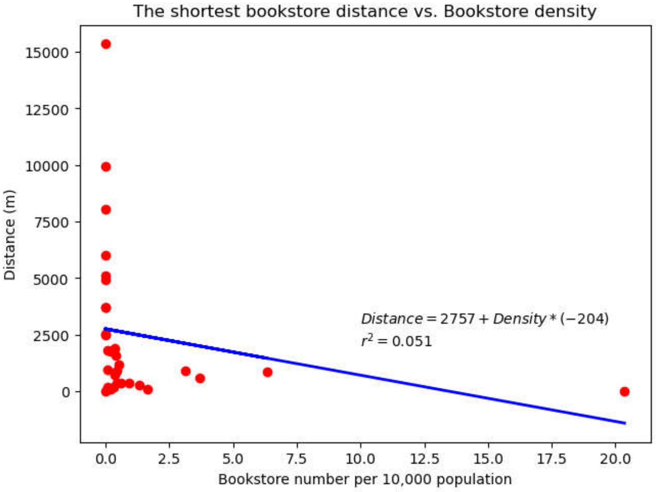
The bar char plots show clearly that in certain local regions the bookstore density were 0 and the closest bookstore were more than 8000 m away.



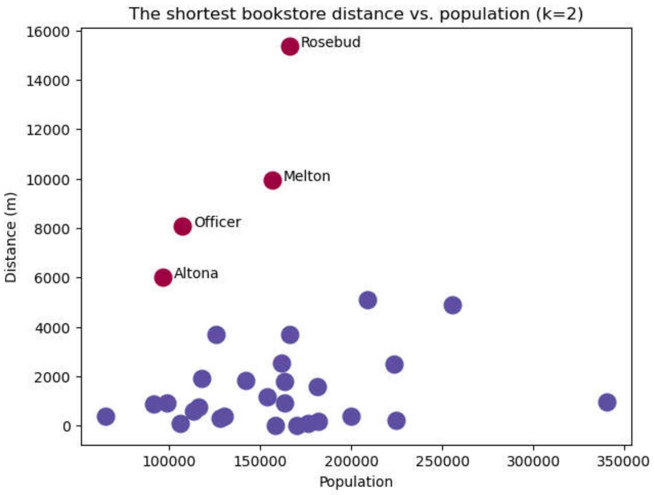


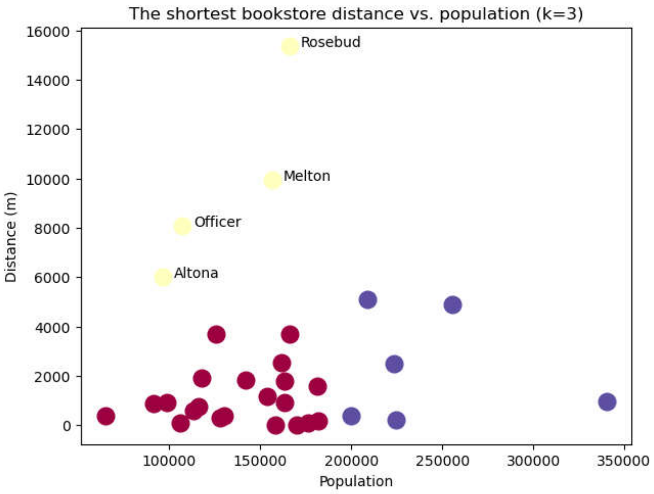
The scatter plots did not show any coorelations between the variables selected, neither did the linear regression. For both scatter plots, the linear regression results had an extremly low r^2 values, 0.0002 and 0.05, respectively.

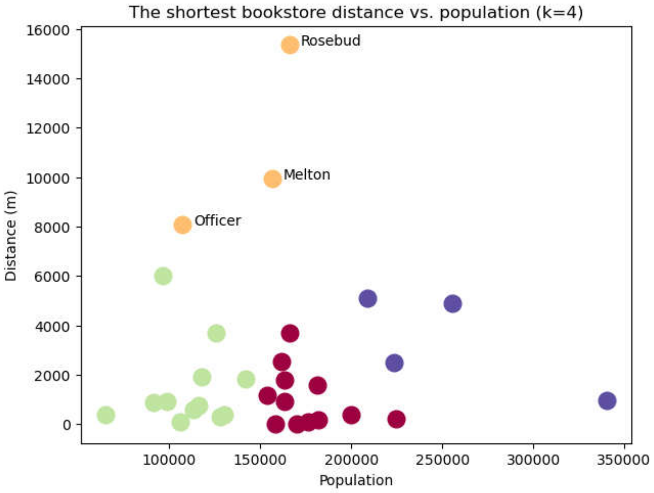


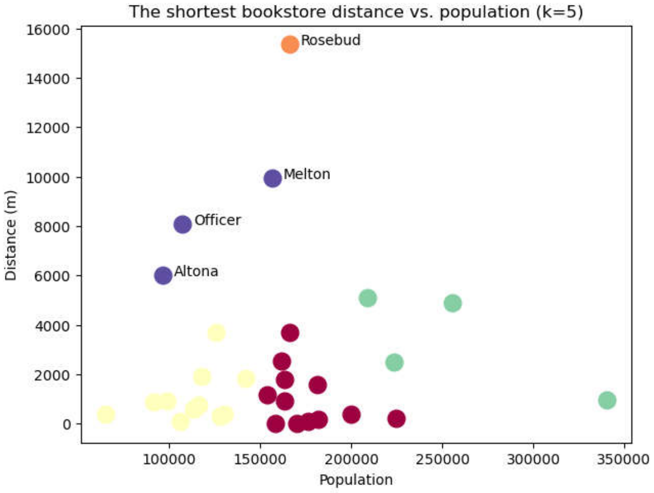


The k-Means clustering analyses show very interesting results. The numbers of clusering ranged from 2 to 5. No matter which k values selected in this range, the few local regions, such as **Rosebud**, **Melton**, **Officer**, and **Altona**, were separated from other local regions. The reason might be that the combination of the distance and population.


The shortest bookstore distance vs. population (k=2)
Distance (m)
Population
0
2000
4000
6000
8000
10000
12000
14000
16000
100000
150000
200000
250000
300000
350000
Rosebud
Melton
Officer
Altona


The shortest bookstore distance vs. population (k=3)
Distance (m)
Population
0
2000
4000
6000
8000
10000
12000
14000
16000
100000
150000
200000
250000
300000
350000
Rosebud
Melton
Officer
Altona


The shortest bookstore distance vs. population (k=4)
Distance (m)
Population
0
2000
4000
6000
8000
10000
12000
14000
16000
100000
150000
200000
250000
300000
350000
Rosebud
Melton
Officer


The shortest bookstore distance vs. population (k=5)
Distance (m)
Population
0
2000
4000
6000
8000
10000
12000
14000
16000
100000
150000
200000
250000
300000
350000
Rosebud
Melton
Officer
Altona

Discussion

Based on the previous study, the local government are of **Rosebud**, **Melton**, **Officer**, and **Altona**, could be selected as the possible locations for a new bookstore. Whether to setup a new one, might be considered with other factors, which were not studied here.

The method used in the current study was relatively simple. There might be several ways to refine the study. For example, in the current study, only the center of the LGA was considered. If the study was finer, a few more locations in an LGA should be considered. Also, the population was not segmentated, while it should be done for a finer study.

Regarding the tools and packages used in the current study, Foursquare and machine learning packages were extremly powerful and should be studied deeply for a finer future study.

```
In [ ]:
```