COURSERA CAPSTONE

BATTLE OF THE NEIGHBORHOODS - SINGAPORE

Introduction / Business Problem

A group of young entrepreneurs would like to start a new cafe business in Singapore. However, they are unsure of where to locate their new cafe. They would their cafe to cater to a wide group of people. They have heard that data science might be able to give them better insights in the location and have given me this project to use data science techniques to assist them.

Data

To approach this data science project, firstly, data regarding the neighbors in Singapore are to be retrieved from Wikipedia (https://en.wikipedia.org/wiki/Planning_Areas_of_Singapore)

A screenshot of the relevant table is shown below:

Name (English)	Malay	Chinese	Pinyin	Tamil	Region	Area (km2)	Population[6]	Density (/km2)
Ang Mo Kio		宏茂桥	Hóng mào qiáo	ஆங் மோ கியோ	North-East	13.94	165,710	12,000
Bedok	*	勿洛	Wù luò	பிடோ	East	21.69	281,300	13,000
Bishan		碧山	Bì shān	பீஷான்	Central	7.62	88,490	12,000
Boon Lay		文礼	Wén lǐ	பூன் லே	West	8.23	30	3.6
Bukit Batok	*	武吉巴督	Wǔjí bā dū	புக்கிட் பாத்தோக்	West	11.13	144,410	13,000
Bukit Merah	*	红山	Hóng shān	புக்கிட் மேரா	Central	14.34	151,870	11,000
Bukit Panjang	*	武吉班让	Wǔjí bān ràng	பக்கிட் பஞ்சாங்	West	8.99	140,820	16,000

Unfortunately, this Wikipedia page does not contain latitude and longitude information of the neighborhoods, which will be filled in later.

Firstly, though, the table from the Wikipedia contains information which are not required, hence columns Malay, Chinese, Pinyin and Tamil are dropped. A screenshot of the partial list of data at this stage is shown below:

[7]:

	Neighborhood	Population	Region	Area(km2)	Density(/km2)
0	Ang Mo Kio	165710	North-East	13.94	12000
1	Bedok	281300	East	21.69	13000
2	Bishan	88490	Central	7.62	12000
3	Boon Lay	30	West	8.23	3.6
4	Bukit Batok	144410	West	11.13	13000
5	Bukit Merah	151870	Central	14.34	11000

So back to the issue of missing latitude and longitude data, using nominatim from geopy.gelocator, the required data is retrieved and added to each neighborhood.

During this process, three of the neighborhoods (Downtown Core, Western Islands and Museum) were not recognized by OpenStreetMap and no latitude and longitude data were returned for these three neighborhoods. Upon further internet search, more detailed data regarding the neighborhood were obtained from the following file from the Department of Statistics Singapore website:
https://www.singstat.gov.sg/-/media/files/find_data/population/statistical_tables/tablea12-2000-2018.xls

From this file, based on the largest population in the subzones of the neighborhood, three other names were chosen (Bugis, Jurong Island and Dhoby Ghaut) to replace the three not recognized by OpenStreetMap.

A screenshot of the partial list of neighborhood with latitude and longitude data added:

[9]:

	Neighborhood	Population	Region	Area(km2)	Density(/km2)	Latitude	Longitude
0	Ang Mo Kio	165710	North-East	13.94	12000	1.369842	103.846609
1	Bedok	281300	East	21.69	13000	1.323976	103.930216
2	Bishan	88490	Central	7.62	12000	1.351455	103.848263
3	Boon Lay	30	West	8.23	3.6	1.345640	103.711802
4	Bukit Batok	144410	West	11.13	13000	1.349057	103.749591
5	Bukit Merah	151870	Central	14.34	11000	1.280628	103.830591
6	Bukit Panjang	140820	West	8.99	16000	1.377921	103.771866

With the latitude and longitude data obtained for all neighborhoods, the latitude and longitude, as well as the name of the neighborhood, are then used to retrieve the location data of the venues (with a radius setting of 1000m) from Foursquare. A total of 2991 venues in 282 unique categories were retrieved.

Screenshot of the partial list of venues::

[26]:

	Neighborhood	Accessories Store	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Arcade	Art Gallery	...	Waterfall	Waterfront
0	Ang Mo Kio	0	0	0	0	0	0	0	0	0	...	0	0
1	Ang Mo Kio	0	0	0	0	0	0	0	0	0	...	0	0
2	Ang Mo Kio	0	0	0	0	0	0	0	0	0	...	0	0
3	Ang Mo Kio	0	0	0	0	0	0	0	0	0	...	0	0
4	Ang Mo Kio	0	0	0	0	0	0	0	0	0	...	0	0

Screenshot of the final dataset's shape:

```
[27]: sg_onehot.shape

[27]: (2991, 282)
```

The dataset is now ready for further analysis (which will be covered in next week's submission)