The Battle of Neighborhoods - Blocks Effective report

Applied Data Science Capstone
IBM Data Science Professional Certificate

Marta Filipa de Oliveira Ferreira das Neves Nabais

July 13, 2020

1 Introduction/Business Problem

"Beauty perishes in life, but is immortal in art"

- Leonardo da Vinci

Blocks Effective is a newly founded start-up architectural firm that prizes itself on catering not only to its clients dream-house design, but also dream location. For this, we make use of the latest machine learning algorithms. In this case report, a young middle-class couple would like to settle around the London area, preferably in a location with a lot of open spaces that would serve as inspiration for the artistry of one of them (a writer/painter) and some sports facilities, to serve as an escape to the fast-paced business life of the other. They are also worried about the current COVID-19 situation and would prefer to be located in an area less affected by the pandemic.

2 Data description

2.1 Data sources

Data were taken from different sources:

- **Wikipedia list of areas in London**, Wikipedia page, containing "Districts of London".
- **Coronavirus (COVID-19) in the UK database**, total and daily UK cases.COVID-19 cases are identified by taking specimens from people and sending these specimens to

laboratories around the UK to be tested. If the test is positive, this is a referred to as a lab-confirmed case.There are separate reporting processes for each of the 4 Nations of the UK. All 4 Nations provide data based on tests carried out in NHS (and PHE) laboratories. These represent 'pillar 1' of the Government's mass testing programme. In addition, testing by commercial partners ('pillar 2' of the mass-testing programme) are now added by the individual Nations before being sent to the the Department for Health and Social Care (DHSC). These are submitted to Public Health England (PHE) to display on the dashboard. The 4 figures are not all taken from the same cut-off time: England and Scotland counts are as at 9am on the day of publication; Wales counts are as at 7am on the day of publication; Northern Ireland counts are from different times on the morning of publication.

- **Foursquare**, a location technology platform dedicated to improving how people move through the real world. I have extracted location information on the type of venues surrounding the California area.

3 Methodology

All analyses were conducted using Jupyter Notebook (Python 3.8.2 kernel).

3.1 Data cleaning

A list of all London districts and neighborhoods was scraped from Wikipedia, using the Python library BeautifulSoup. Data was parsed into a pandas dataframe, with 4 columns (Neighborhood, Borough, Town and PostCode). Initial data cleaning consisted of several steps:

- Expanding the dataframe so neighborhoods with more than one postcode were separated by rows
- Retaining postcodes belonging to LONDON town
- Removing the "Town" column after cleaning the data
- Getting coordinates for each postcode, using the python library pgeocode

3.2 Exploratory Data Analysis

I first visualized all neighborhoods in London geographically, using the folium Python library. Then I focused on one of the postcodes (W3, for Acton) to verify if the Foursquare API was working properly. I used the Forsquare API to search for nearby venues located within 500 meter radius of the W3 postcode (Acton). I have limited my search to 100 venues, maxmimum. Then I repeated this search for all the neighborhoods in London. To get a summary for each neighborhood, I used one-hot encoding to get dummy variables for all venue types, in each neighborhood. I then calculated the frequency of each venue type, within in each neighborhood and calculated the top 10 most common venues within each neighborhood.

3.3 Unsupervised K-means clustering

I divided the data into non-overlapping clusters, based on similarity between groups, using K-means algorithm (unsupervised clustering method), with the python library sklearn. I have set the number of clusters to 5 and then visualized the clusters using the folium Python library.

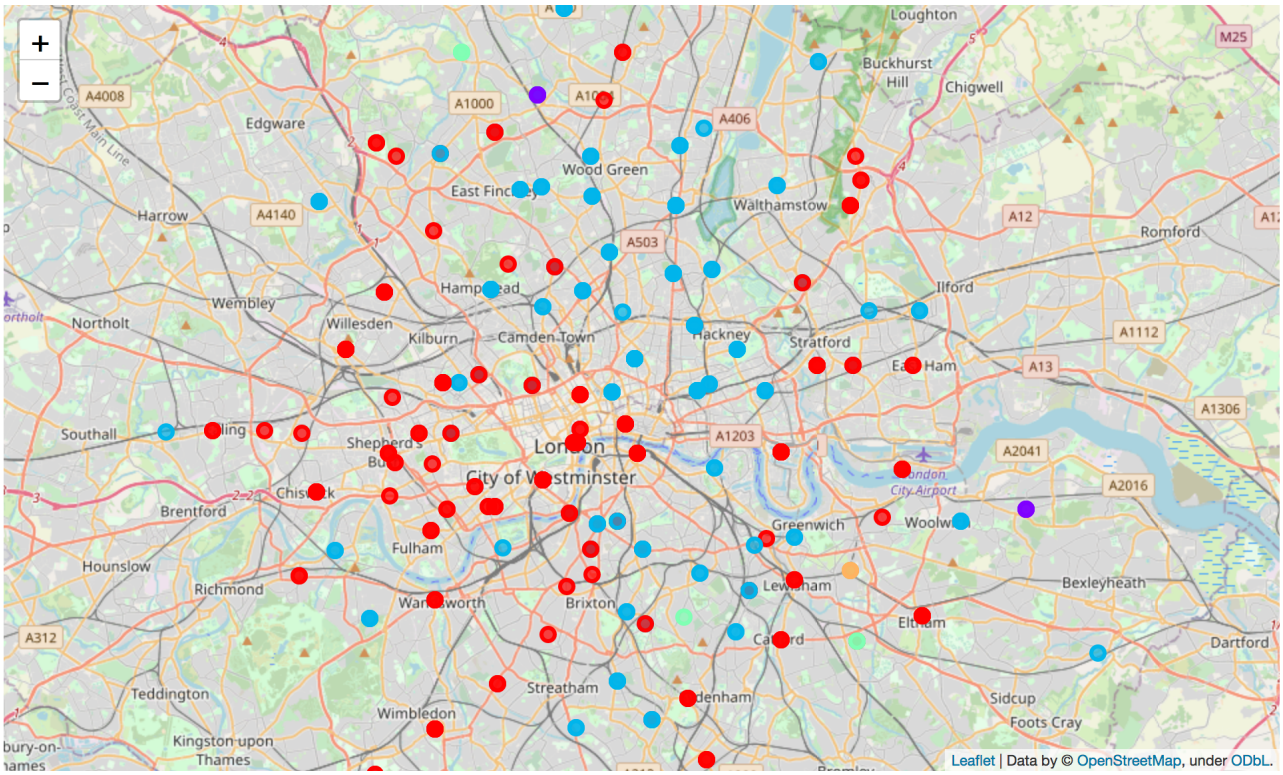
3.4 COVID-19 risk assessment

Finally, I assessed the the number of COVID-19 current infection rate (over 14-days periods) over the London area, using latest coronavirus update, from the UK government.

4 Results & Discussion

4.1 Determining the best neighborhood with K-means clustering

The results from K-means clustering show we can categorize the London area neighborhoods into 5 clusters based on the type and frequency of venues located in each neighborhood. The results can be visualized in Figure 4.1, below.



- Cluster 1: Neighborhoods with many pubs, coffee shops and restaurants
- Cluster 2: Neighborhoods with many grocery stores
- Cluster 3: Neighborhoods with many pubs and Asian restaurants
- Cluster 4: Neighborhoods with parks and transport services
- Cluster 5: Neighborhoods with many transportation services

Based on the clients preferences, I have determined cluster number 4 as the best potential neighborhoods for them to settle (plenty of open spaces and parks), as in Table 4.1, below.

Out[30]:

	Neighborhood	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
51	Blackheath Royal Standard	0.017500	3.0	Park	Transportation Service	Bus Stop	Fast Food Restaurant	Pub	Deli / Bodega	Ethiopian Restaurant	Food Court	Food & Drink Shop	Flower Shop
51	Blackheath Royal Standard	0.020900	3.0	Park	Transportation Service	Bus Stop	Fast Food Restaurant	Pub	Deli / Bodega	Ethiopian Restaurant	Food Court	Food & Drink Shop	Flower Shop
96	Chinbrook	0.020900	3.0	Pub	Bus Stop	Fast Food Restaurant	Park	Flower Shop	Flea Market	Fish Market	Food & Drink Shop	Fish & Chips Shop	Film Studio
147	East Dulwich	-0.068767	3.0	Pub	Park	Grocery Store	Bus Stop	Cycle Studio	Fast Food Restaurant	Ethiopian Restaurant	Event Space	Exhibit	Falafel Restaurant
249	Horn Park	0.020900	3.0	Pub	Bus Stop	Fast Food Restaurant	Park	Flower Shop	Flea Market	Fish Market	Food & Drink Shop	Fish & Chips Shop	Film Studio
281	Lee	0.020900	3.0	Pub	Bus Stop	Fast Food Restaurant	Park	Flower Shop	Flea Market	Fish Market	Food & Drink Shop	Fish & Chips Shop	Film Studio
347	Oakleigh Park	-0.183320	3.0	Park	Golf Course	Metro Station	English Restaurant	Ethiopian Restaurant	Event Space	Exhibit	Falafel Restaurant	Farm	Yoga Studio
468	Totteridge	-0.183320	3.0	Park	Golf Course	Metro Station	English Restaurant	Ethiopian Restaurant	Event Space	Exhibit	Falafel Restaurant	Farm	Yoga Studio
514	Whetstone	-0.183320	3.0	Park	Golf Course	Metro Station	English Restaurant	Ethiopian Restaurant	Event Space	Exhibit	Falafel Restaurant	Farm	Yoga Studio

Finally, to pinpoint the best neighborhood I have pulled the current infection rate of COVID-19 the boroughs containing the neighborhoods in cluster 4 (Figure 4.1, with the Greenwhich or Lewisham boroughs being the least affected.

In [37]:
```
greenwich_covid_rate = int(corona_df[corona_df['Borough'] == 'Greenwich'].sort_values(['DailyTotalLabConfirmedCasesRate', 'Dat
e'], ascending = False).drop_duplicates('Borough', keep = 'first')['DailyTotalLabConfirmedCasesRate'])
lewisham_covid_rate = int(corona_df[corona_df['Borough'] == 'Lewisham'].sort_values(['DailyTotalLabConfirmedCasesRate', 'Dat
e'], ascending = False).drop_duplicates('Borough', keep = 'first')['DailyTotalLabConfirmedCasesRate'])
southwark_covid_rate = int(corona_df[corona_df['Borough'] == 'Southwark'].sort_values(['DailyTotalLabConfirmedCasesRate', 'Dat
e'], ascending = False).drop_duplicates('Borough', keep = 'first')['DailyTotalLabConfirmedCasesRate'])
barnet_covid_rate = int(corona_df[corona_df['Borough'] == 'Barnet'].sort_values(['DailyTotalLabConfirmedCasesRate', 'Date'], a
scending = False).drop_duplicates('Borough', keep = 'first')['DailyTotalLabConfirmedCasesRate'])

print('The current daily rate of COVID-19 infections in Greenwich is {}'.format(greenwich_covid_rate))
print('The current daily rate of COVID-19 infections in Lewisham is {}'.format(lewisham_covid_rate))
print('The current daily rate of COVID-19 infections in Southwark is {}'.format(southwark_covid_rate))
print('The current daily rate of COVID-19 infections in Barnet is {}'.format(barnet_covid_rate))
```

```
The current daily rate of COVID-19 infections in Greenwich is 334
The current daily rate of COVID-19 infections in Lewisham is 395
The current daily rate of COVID-19 infections in Southwark is 456
The current daily rate of COVID-19 infections in Barnet is 408
```

Based on these results, I have recommended the couple to settle on the Blackheath Royal Standard neighborhood, as in Table 4.1.

Out[38]:

	Neighborhood	Borough	PostCode	Latitude	Longitude
51	Blackheath Royal Standard	Greenwich	SE3	51.4672	0.017500
51	Blackheath Royal Standard	Greenwich	SE12	51.4448	0.020900
96	Chinbrook	Lewisham	SE12	51.4448	0.020900
147	East Dulwich	Southwark	SE22	51.4521	-0.068767
249	Horn Park	Greenwich, Lewisham	SE12	51.4448	0.020900
281	Lee	Lewisham	SE12	51.4448	0.020900
347	Oakleigh Park	Barnet	N20	51.6333	-0.183320
468	Totteridge	Barnet	N20	51.6333	-0.183320
514	Whetstone	Barnet	N20	51.6333	-0.183320

5 Conclusions

K-means clustering is an effective method to categorize neighborhoods in London according to the similarity of venue types. This strategy has allowed me to best advise the clients to settle on Blackheath Royal Standard, according to their preferences.