Applied Data Science Capstone

Final Report

1/1/2021

Walid Ismail

Contents

[Introduction / Business Problem 3](#_Toc60327497)

[Data 3](#_Toc60327498)

[Methodology 4](#_Toc60327499)

[Results 4](#_Toc60327500)

[Discussion 4](#_Toc60327501)

[Conclusion 4](#_Toc60327502)

[References 4](#_Toc60327503)

# Introduction / Business Problem

London is a big European city with hundreds of venues spread all over its many neighborhoods and competition is hard, so identifying locations that provide an opportunity for a profitable business and where there is real customer need is paramount.

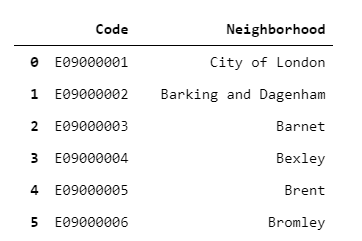
Our client is planning to open a restaurant / franchise in London and is seeking our advice to help them identify suitable location(s) where they can establish their business in the busy and condensed city of London, UK.

Our analysis will aim to help them with this decision.

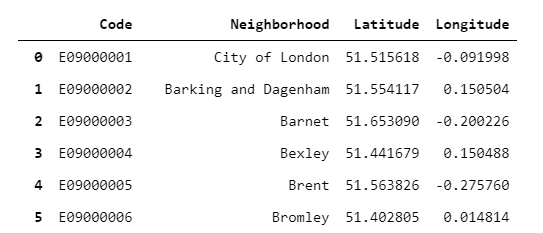
# Data

The data used for our analysis was obtained from two main sources:

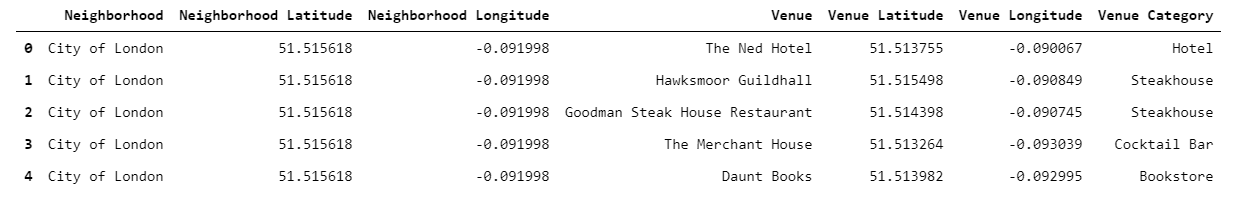
The first one is London, UK e-gov data store website at the following link: [London Borough Profiles and Atlas - London Datastore](https://data.london.gov.uk/dataset/london-borough-profiles). This data includes the list of London 33 boroughs (including City of London borough) as shown in the excerpt figure below:



The names of London boroughs were then supplemented with the co-ordinates for each borough. The latitude and longitude information for each borough were obtained by querying The [Geopy](https://geopy.readthedocs.io/en/stable/) library and appended to the name of each London shown as shown in figure 2.



The second source of information and data is the [Foursquare.com](https://foursquare.com/developers/) location data database. This online database was used to obtain a list of 100 popular venues within 1000 meters from the center of each borough. This data is available publicly free-of-charge via Foursquare API calls. The location data for venue includes venue name, latitude, longitude and Category. A sample of the data is shown below:

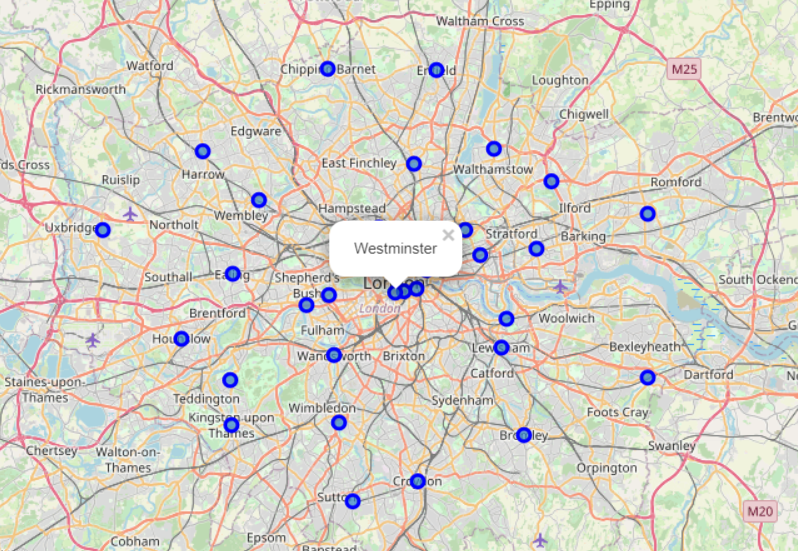


# Methodology

Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learning were used and why.

Our analysis centered on clustering the neighborhood of London in order to find groups similar neighborhoods and discover their main characteristics and obtain insights about each cluster of neighborhoods that will guide our recommendation on the best neighborhoods to open a new restaurant in London.

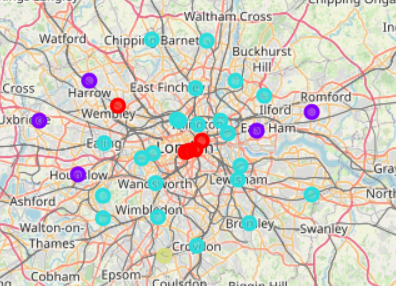
The first step in our analysis was to prepare the data for discovery and exploration through foursquare location data. This step included loading and verifying London neighborhoods, fetch latitude and longitude information for each neighborhood. A folium map of the 33 boroughs is shown below:



The second step is exploring each neighborhood venue. We used foursquare REST API calls to obtain 100 popular venues for each neighborhood within a 1000m radius from each neighborhood center. Our exploration revealed 2055 acorss the 33 boroughs spread across 252 categories.

The third step was to one-hot encode the categories in order to have a matrix of boroughs and their corresponding categories and use this matrix for clustering using k-means algorithm.

Next we sorted each borough’s list of categories of venue in descending order to reveal the 10 most popular categories of venues in each borough. We followed that by clustering the neighborhoods into four clusters using our 252 categories as attributes/features. The resulting clusters are shown below:



The clustering shows the following interesting information:

* The first cluster is composed of 4 main borough at the center of London in addition to one similar borough at the outskirt of the city, namely:
  + City of London
  + Westminster
  + Lambeth
  + Southwark
  + Brent
* The second cluster consists of 22 boroughs that encapsulate the center of the city.
* The third cluster includes 5 boroughs at the edge of the city.
* The fourth cluster includes one borough only, it is Sutton.

Finally, we analyzed each cluster using then 10 popular categories of venues for boroughs of the cluster

# Results

Results section where you discuss the results.

# Discussion

Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.

# Conclusion

Conclusion section where you conclude the report.

# References

* Geopy Library: <https://geopy.readthedocs.io/en/stable/>
* Foursquare.com API: <https://foursquare.com/developers/>