For the second week, the final deliverables of the project will be:

1. A link to your Notebook on your Github repository, showing your code. (15 marks)
2. A full report consisting of all of the following components (15 marks):

* Introduction where you discuss the business problem and who would be interested in this project.
* Data where you describe the data that will be used to solve the problem and the source of the data.
* 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 learnings were used and why.
* Results section where you discuss the results.
* Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.
* Conclusion section where you conclude the report.

3. Your choice of a presentation or blogpost. (10 marks)

**Title: Tourism enhancers index by neighbourhoods in London**

**Introduction**

London is one of the most popular tourist destinations in the whole world, join Paris and Bangkok. Every year London city attracts millions of visitors from other countries. During 2019 London received 19.09 millions of tourists according to the Mastercard’s annual Global Cities Index publishing. The city attracted 20.42 million international visitors in 2018, making it one of the world's most visited in terms of international visits.[[1]](https://en.m.wikipedia.org/wiki/Tourism_in_London" \l "cite_note-1) It welcomed an additional 27.8 million overnighting domestic tourists in 2017[[2]](https://en.m.wikipedia.org/wiki/Tourism_in_London" \l "cite_note-2) and had 280 million daytrippers in 2015 [[3]](https://en.m.wikipedia.org/wiki/Tourism_in_London" \l "cite_note-3). Overall London sees in nearly 50 million overnighters each year, and over 300 million visitors if including daytrippers.

The tourism industry is one of the engines of economic growth in the UK. London first works int the Tourism Sector Deal proposal. This proposal defends and enhances London's role as a gateway, open and welcoming destination for both business and leisure visitors [<https://www.londonfirst.co.uk/what-we-do/economy-and-tax/tourism>].

Tourism agencies published the statistics related to the most touristic visited areas. In 2018 Hangluggageonly published the 15 London most touristic visited areas: Southbank, Bermondsey Street, Camden, Notting Hill, Peckham, Whitechapel, Kings Cross, Fitzrovia, Covent Garden, Elephant and Castle, City of London, Richmond, Greenwich, Shoreditch, Westminster.[<https://handluggageonly.co.uk/2018/10/24/15-of-the-best-areas-in-london-you-have-to-visit/>] Tripadvisor published the 10 best neighbourhoods to explore in London: Mayfair, Shoreditch, Chelsea, Greenwich, Southwark, Brixton, Notting Hill, Camden, Soho, Kensington. **[https://**[**www.tripsavvy.com/best-neighbourhoods-to-explore-**](http://www.tripsavvy.com/best-neighbourhoods-to-explore-) **in-london-4129386].**

Traditionally the tourism market and growth has been analyzed with studies based on coarse granularity. New technologies allow segment, find out and develop new potential touristic zones. The resulting indicators belong to different tourist targets independence of the sector needs: retail, e-commerce, health, personal services, hospitality, transport, accommodation, etc.. all applicable as explanatory variables in a geomarketing models.

In this project, motivated by the challenge of understanding urban environments, and based on the opportunities created by geo-enabled social data, we address the problem of comparing boroughs and neighbourhoods in London city.The problem we study has applications for recommending locations in London city as tourist enhancers. Imagine an entrepreneur looking for areas with tourist profile for opening a new establishment related to the tourist target, for example, where should I open a basement to rent bicycles?: the methods developed here allow matching each neighbourhood in a borough with the most similar neighbourhood in another borough, or any other borough that someone wishes to compare.

[Michael John Hebbert](https://www.britannica.com/contributor/Michael-John-Hebbert/4181), Blake Ehrlich, Hugh D. Clout. London.National Capital. United Kingdom. <https://www.britannica.com/place/London>] Aug 28,2019.

1.Simon Kyte (May 2012). [*Tourism in London*](https://web.archive.org/web/20130228094422/http://www.london.gov.uk/sites/default/files/wp53.pdf) (PDF). [Greater London Authority](https://en.m.wikipedia.org/wiki/Greater_London_Authority). p. 11. [ISBN](https://en.m.wikipedia.org/wiki/International_Standard_Book_Number) [978-1-84781-496-8](https://en.m.wikipedia.org/wiki/Special:BookSources/978-1-84781-496-8). Archived from [the original](http://www.london.gov.uk/sites/default/files/wp53.pdf) (PDF) on 28 February 2013. Retrieved 10 March 2013. Italic or bold markup not allowed in: |publisher= ([help](https://en.m.wikipedia.org/wiki/Help:CS1_errors" \l "apostrophe_markup))

2. ["Domestic overnight stays in London 2008-2017 | UK Statistic"](https://www.statista.com/statistics/487505/nights-spent-by-domestic-visitors-london-united-kingdom/). *Statista*. Retrieved 28 March 2019.

3. ["London and Domestic Tourism Report 2015"](https://webcache.googleusercontent.com/search?q=cache:DgM2p3DVFLIJ:https://www.visitbritain.org/sites/default/files/vb-corporate/Documents-Library/documents/England-documents/london_2015.pdf+&cd=11&hl=en&ct=clnk&gl=uk&client=firefox-b-d). [*https://files.londonandpartners.com/l-and-p/assets/our-insight-london-tourism-review-2014-15.pdf*](https://files.londonandpartners.com/l-and-p/assets/our-insight-london-tourism-review-2014-15.pdf). External link in |website= ([help](https://en.m.wikipedia.org/wiki/Help:CS1_errors" \l "param_has_ext_link))

http://files.londonandpartners.com/l-and-p/assets/london\_tourism\_vision\_aug\_2017.pdf

4. [<https://en.m.wikipedia.org/wiki/Tourism_in_London>]

Description of the problem:

Analyze the segmentation and clustering of neighbourhoods in London

based on the geopolitical distribution, the tourism and the technologic industries.

Analyze which neighbourhoods from New York, USA and London, UK are similars based on the above information

The question is to compare the neighbourhoods of the two cities and determine how similar or dissimilar they are.

Steps:

Segmentation and Clustering of neighbourhood in London

london-borough-profiles.csv

[https://data.london.gov.uk › dataset › london-borough-profiles](https://data.london.gov.uk/dataset/london-borough-profiles) contains

code

Area name

kind:Inner/Outer

GLA population Estimate 2017

Gla Household Estimate 2017

<https://data.london.gov.uk/dataset/gla-population-projections-custom-age-tables>

Geographic variable : oslaua str9 local authority district

<https://www.understandingsociety.ac.uk/sites/default/files/downloads/documentation/data-linkage/mainstage-geographical-lookup-tables.pdf>

<https://latitude.to/articles-by-country/gb/united-kingdom/4105/list-of-london-boroughs>

Segmentation and Clustering of neighbourhood with Tourism information

Segmentation and Clustering of neighbourhood with Tech information

Segmentation and Clustering of neighbourhood of the two cities New York and London

Discussion of the background:

Description of the used data:

Solution of the problem:

**Title: Tourism enhancers index by neighbourhoods in London**

London is one of the most popular tourist destinations in the whole world, join Paris and Bangkok. Every year London city attracts millions of visitors from other countries. During 2019 London received 19.09 millions of tourists according to the Mastercard’s annual Global Cities Index publishing.

The tourism industry is one of the engines of economic growth in the UK. London first works int the Tourism Sector Deal proposal. This proposal defends and enhances London's role as a gateway, open and welcoming destination for both business and leisure visitors [<https://www.londonfirst.co.uk/what-we-do/economy-and-tax/tourism>].

Tourism agencies published the statistics related to the most touristic visited areas. In 2018 Hangluggageonly published the 15 London most touristic visited areas: Southbank, Bermondsey Street, Camden, Notting Hill, Peckham, Whitechapel, Kings Cross, Fitzrovia, Covent Garden, Elephant and Castle, City of London, Richmond, Greenwich, Shoreditch, Westminster.[<https://handluggageonly.co.uk/2018/10/24/15-of-the-best-areas-in-london-you-have-to-visit/>] Tripadvisor published the 10 best neighbourhoods to explore in London: Mayfair, Shoreditch, Chelsea, Greenwich, Southwark, Brixton, Notting Hill, Camden, Soho, Kensington. **[https://**[**www.tripsavvy.com/best-neighbourhoods-to-explore-**](http://www.tripsavvy.com/best-neighbourhoods-to-explore-) **in-london-4129386].**

Traditionally the tourism market and growth has been analyzed with studies based on coarse granularity. New technologies allow segment, find out and develop new potential touristic zones. The resulting indicators belong to different tourist targets independence of the sector needs: retail, e-commerce, health, personal services, hospitality, transport, accommodation, etc.. all applicable as explanatory variables in a geomarketing models.

In this project, motivated by the challenge of understanding urban environments, and based on the opportunities created by geo-enabled social data, we address the problem of comparing boroughs and neighbourhoods in London city.The problem we study has applications for recommending locations in London city as tourist enhancers. Imagine an entrepreneur looking for areas with tourist profile for opening a new establishment related to the tourist target, for example, where should I open a basement to rent bicycles?: the methods developed here allow matching each neighbourhood in a borough with the most similar neighbourhood in another borough, or any other borough that someone wishes to compare.

Mathematical Problem Formulation

Let’s the following borough-neighbourhood search problem:

We are giving the London city and its sets of boroughs B={ B1, B2,…., Bm} and neighbourhoods N={N1.N2….., Np}, and a neighbourhood Ni in a borough Bk and a subset of target boroughs B’ of B. Our goal is to find a neighbourhood Nj in some borough Bs of B’ such that the distance d(Ni, Nj) is minimized.

We address the special case B’=B, which means searching for the most similar neighbourhood in all boroughs.

Problem Background

There are a wide literature relative to the formulation and solution of this kind of mathematical problems that can be resumed in the study of a global optimization problem on the space (human actions, user, locations, temporal coordinate, venue features) and its research activity is active in the looking for computational algorithms more efficient. Relevant information relative finding similarities between neighbourhoods or cities can be found in:

1)-Geraud Le Falher. Finding similar neighbourhoods across cities by mining human urban activity. Where is Beverly Hills in your town?. Master’s Thesis Espoo, August 9, 2014. Supervisor: Professor Aristides GionisAdvisor: Michael Mathioudakis, PhD. Aal to University. School of Science[. http://research.ics.aalto.fi/dmg/](http://research.ics.aalto.fi/dmg/)

2)-Geraud Le Falher, Aristides Gionisand Michael Mathioudakis. Where is the Soho of Rome?. Measures and Algorithms for Finding Similar Neighbourhoods in Cities. Proceeding of Ninth International AAAI Conference on Web and Social Media.Oxford, England, May 26–29, 2015.

3)-Daniel Preotiuc Pietro, Justin Cranshaw and TaeYano. Exploring venue-based city-to-city similarity measures. 2013/08/11, DO - 10.1145/2505821.2505832

4)Simon Pemberton and Rachel Humphris. Locality, neighbourhood and health: a literature review. Irisworking Paper Series,NO.13/2016www.birmingham.ac.uk/iris

Theoretical Method Background

There are many methods to solve the neighbourhood similarity-search problem formulated above. In all cases, the methods are related to clustering and learning metric. In this preliminary investigation, we applying the popular clustering method named k-means algorithm [J. MacQueen.Some methods for classification and analysis of multivariate observations.1967]. This method belongs to unsupervised methods class and it allows to partition a given dataset into several groups on some notion of similarity. It is an iterative algorithm based on the update the positions of the centroids from the initial selection of centroids. This process converges when the position of centroids doesn’t change between two iterations. In practice, the algorithm is fast but only guarantees a local optimum of the sum of squares of distance each data to the centroid within a cluster. An important drawback of this method is that the number of centroids has to be specified at the beginning. Also, the resulting clusters are linearly separable, which may not be true in the original dataset.

**Data and its description**

Our dataset consists of the following public datasets:

1-Visitors-London-country-year [London Datastore]

<https://data.london.gov.uk/dataset/number-international-visitors-london>

2-London-borough-profiles [London Datastore]

https://data.london.gov.uk/dataset/london-borough-profiles

3-neighbourhoods, boroughs-coordinates.kml and areas-coordinates.kml

https://en.wikipedia.org/wiki/List\_of\_areas\_of\_London

4-Foursquare

**London's** defining **characteristic** is an absence of overall form. It is physically a polycentric city, with many core districts and no clear hierarchy among them.

-London Boroughs

[https://en.wikipedia.org/wiki/London\_boroughs]

London has 32 boroughs that corresponding to the local authority that makeup Geater London and each one of them is governed by a London borough council. They were created at the same time as Greater London on 1 April 1965 by the London Government Act 1963.

From this time, it has twelve inner boroughs and twenty outer boroughs.

The London boroughs have populations of around 150,000 to 300,000. Inner London boroughs tend to be smaller, in both population and area, and more densely populated than Outer London boroughs.

We revised different data sources to know the precise number of boroughs and neighbourhoods in London. From the London Datastore repository, we get London-borough-profiles file that contains information related to each borough: Area Code, name, demographic data, etc. From Wikipedia page, we got a list of areas of London that was filtered to get a good correlation between list of boroughs and neighbourhoods. After, we scrapped the Wikipedia page using fastkml, and BeautifulSoup to get the location coordinates of boroughs and neighbourhoods. After filtering and completing the location coordinates list using the geopy library, we got both tables related to boroughs and neighbourhoods with its locations. For our study, we have 33 boroughs and 561 neighbourhoods.

We consider a neighbourhood as an ensemble of services that it offers: parks, bars, restaurants, shops, schools, hotels, transport, arts and entertainment, etc. In the actuality, location-based social systems as Foursquare solves feasibly the difficult task of collecting detailed data about these places. This app prompts their users to tag the locations they check-in to with descriptive categorical labels.

Exploring geotagged activity logs from Foursquare, we gather the description and categorizations of the venues in London city. For the venues, we collected data with the categorical descriptions primary and secondary. Primary category is the most general. For example, “Restaurant” is the primary category, and its derived one is “Indian Restaurant” are typical categories [ This distinction has been used by Daniel Preotiuc et al 2013].

In this work, we take advantage of this feature for our preliminary study. The venue information is easily accessible through a public API, and all venues are annotated with categories of different granularities. We collected Foursquare venues for 33 boroughs and 561 neighbourhoods across the London city, resulting in 1597 and 12083 venues respectively.

Our approach is briefly described as follows. Using geo-enabled data from social media platforms, we represent each venue with a feature vector, that accurately describes the characteristics and the overall activity of the venue. We then devise similarity measures between neighbourhoods, as well as between boroughs, i.e., sets of venues that are geographically close to each other. We address these two problems applying unsupervised K-means clustering method to set of primary and secondary categories separately, and then, we exploit the geo-location relation borough-neighbourhood to conclude the actual similarity between neighbourhoods.

Along with this project, we compute a new dataset with the frequency media of micro territory business by neighbourhoods and the target of the areas most visited in London during 2018 as favourites. We apply the logit method to classification in favourites and non-favourites neighbourhoods.

We apply this project to predict tourist neighbourhoods enhancers that are hidden in the tourist network and join to them show the tourist micro territory Index. This preliminary study has some subjective component. Motivated by this observation, we consider that our study can be enriched by many other types of local data, such as local visitors numbers, accommodation, transportation, weather, air quality, energy consumption, etc.