Capstone Project - The Battle of Neighborhoods (Week 2)

IBM Data Science
Professional Certificate
August 2020

Introduction

- Objective: create a way for a Real State Agency to select the best region of Greater London to offer their clients
- The idea is to find out, by leveraging location data from the Foursquare db (among other), about the regions with the highest concentration of each category of venue.
- Then the Agency can offer flats in regions that match the clients profile. For example, if the favourite form of leisure for one particular client is dining out, the agency can offer flats in the regions with the highest density of restaurants.

Data (1)

- Data will be collected from :
  - Foursquare Places API (https://enterprise.foursquare.com/products/places)
  - Bing Maps REST Services (https://docs.microsoft.com/en-us/bingmaps/rest-services/)
  - Postcode.io (http://api.postcodes.io/)
  - Wikipedia (https://en.wikipedia.org/wiki/London_boroughs)

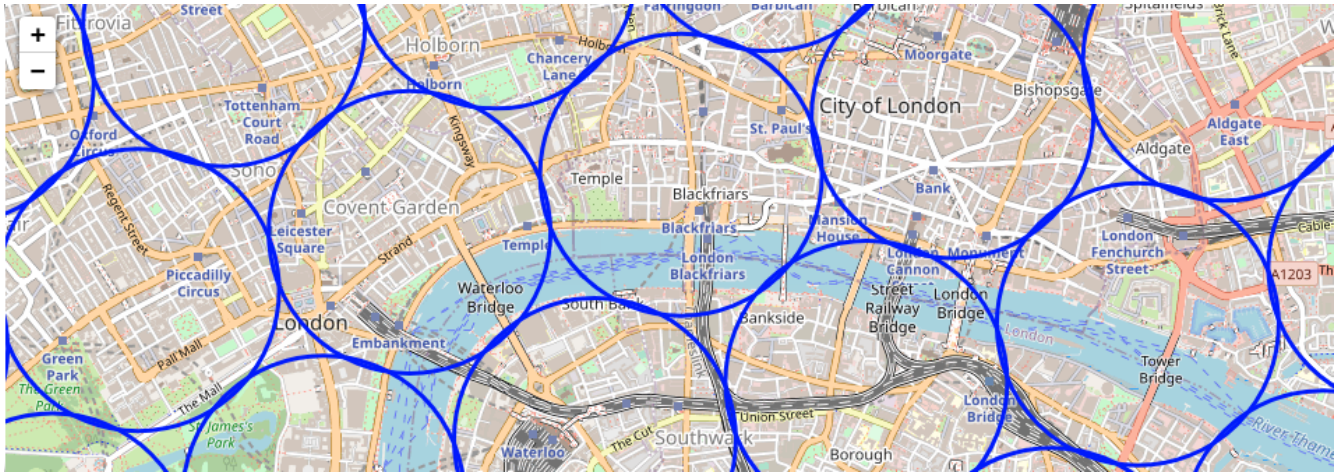
Data (2)

- Foursquare Places API data will be used to obtain the venues.

- Bing Maps REST Services, Postcode.io and Wikipedia will be used to determine if the venues are really located in one of the 32 Greater London boroughs.

Getting and cleaning the data (1)

- Foursquare limits the number of returned venues to 50. To be able to get all venues in Greater London, we divided the area in small circular regions with a radius of 600 meters and queried each one independently, once for each venue category.



Getting and cleaning the data (2)

- We took the latitude and longitude of each venue collected with the “explore“ Foursquare command. Then we used these coordinates to query Postcode.io or Bing to find the borough in which the venue is located.
- The obtained borough was compared to the list of 32 boroughs that are part of Greater London. If the venue was not located in one of these boroughs, it was dropped.

Methodology (1)

- After collecting all venues, grouped by category, we created 2D histograms for each category.
- Histograms were created with 1 km by 1 km bins, after converting the venues‘ coordinates from latitude and longitude to cartesian coordinates (using module pyproj - https://pypi.org/project/pyproj/)
- Bins with counts larger than 80 % of the densest bin were considered high density bins.

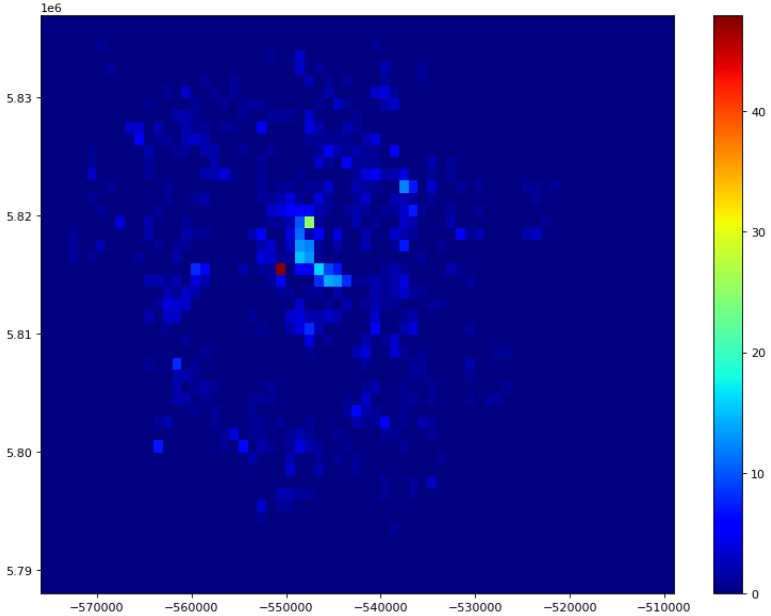
Methodology (2)

- These are the number of venues in each Foursquare category found within the 32 boroughs of Greater London:

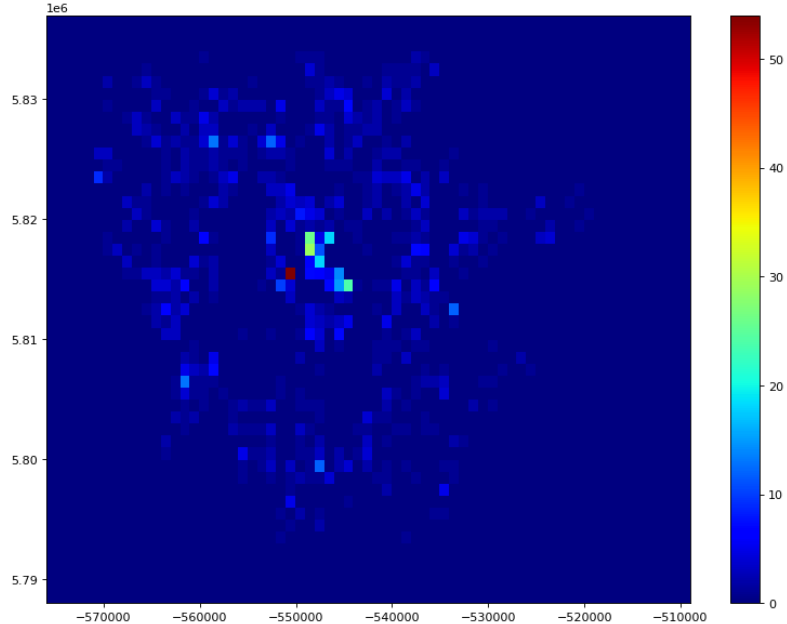
Arts & Entertainment	1029	Outodoor & Recreation	2586
College & University	1322	Professional & Other Places	4326
Event	16	Residence	787
Food	5256	Shop & Service	6234
Nighlife Spot	2446	Travel & Transport	2733

- We dropped the “Event“, as it contains too few venues.

2D Histograms (1)

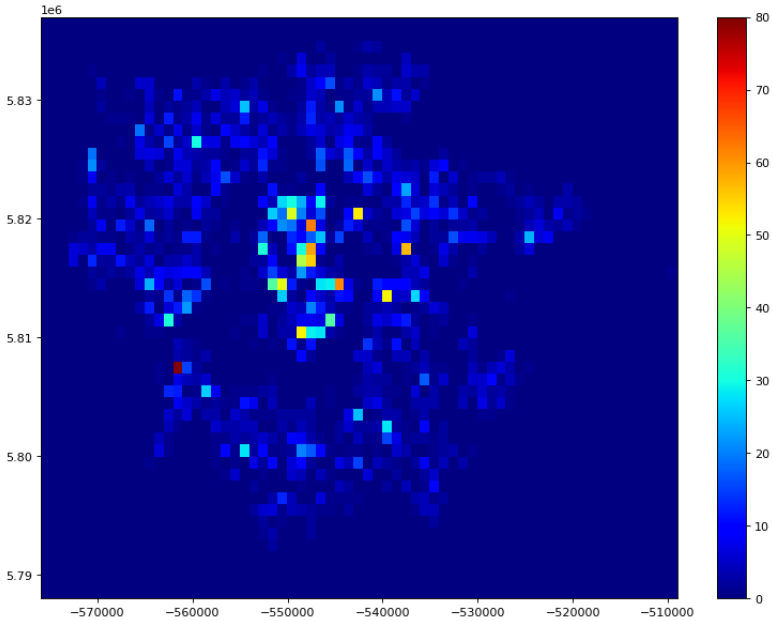


Arts & Entertainment

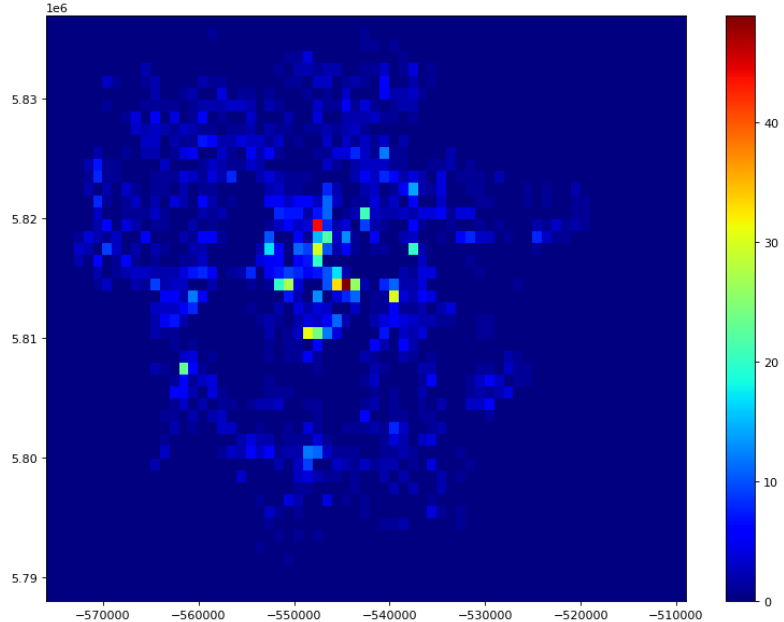


College & University

2D Histograms (2)

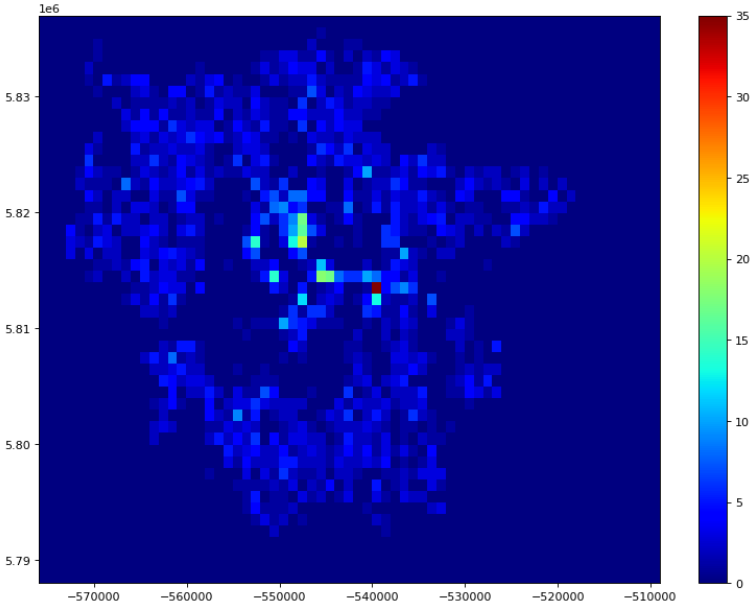


Food

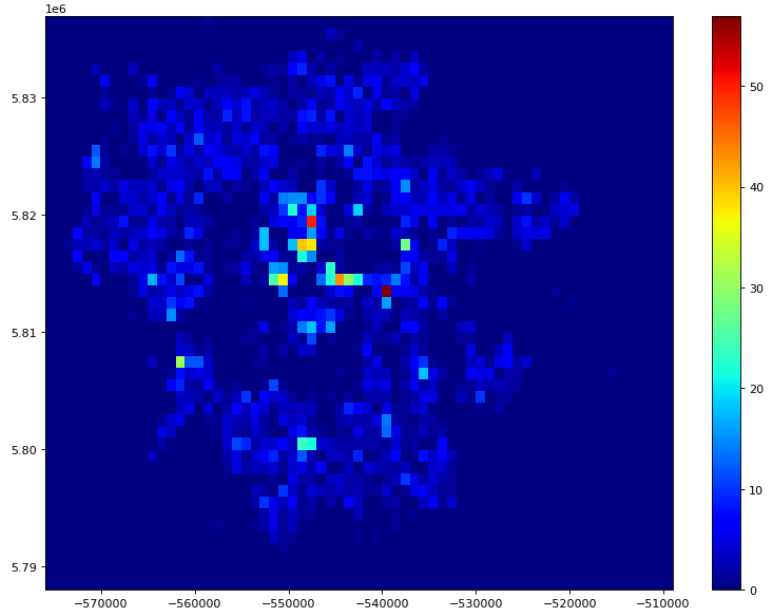


Nightlife Spot

2D Histograms (3)

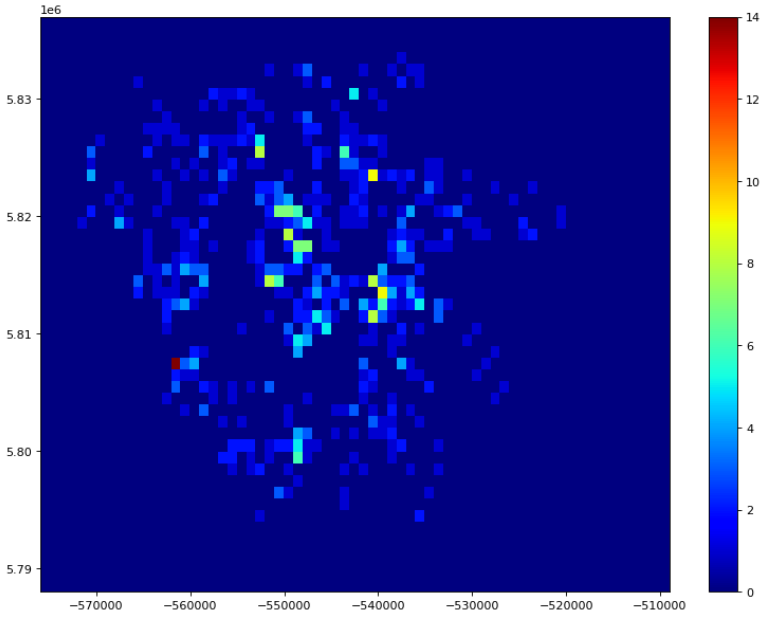


Outdoors & Recreation

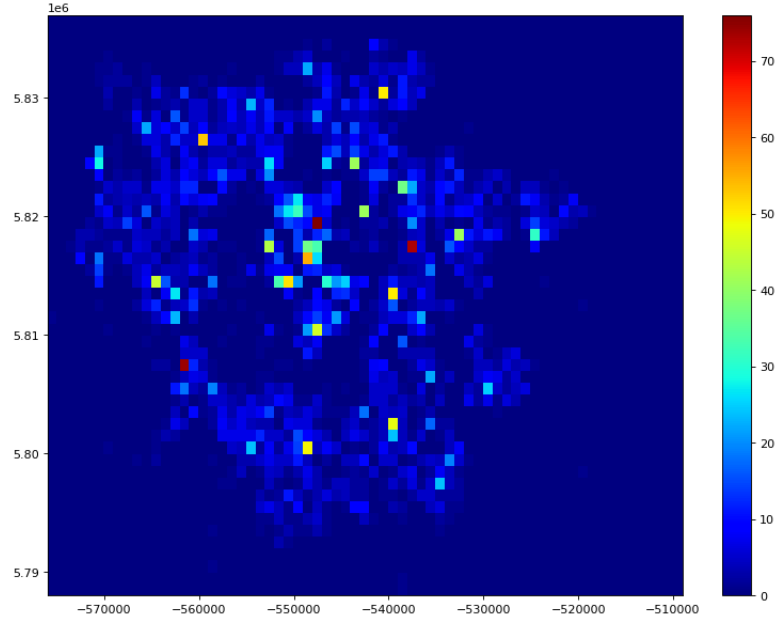


Professional & Other Places

2D Histograms (4)

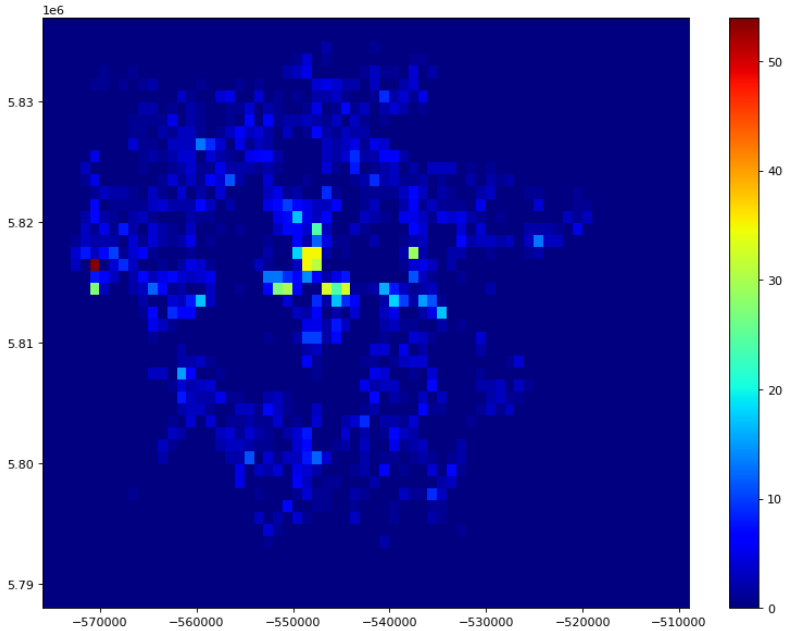


Residence



Shop & Service

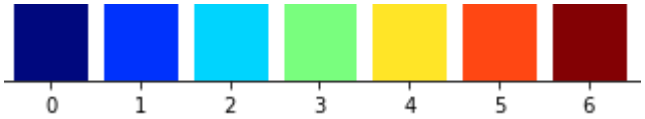
2D Histograms (5)

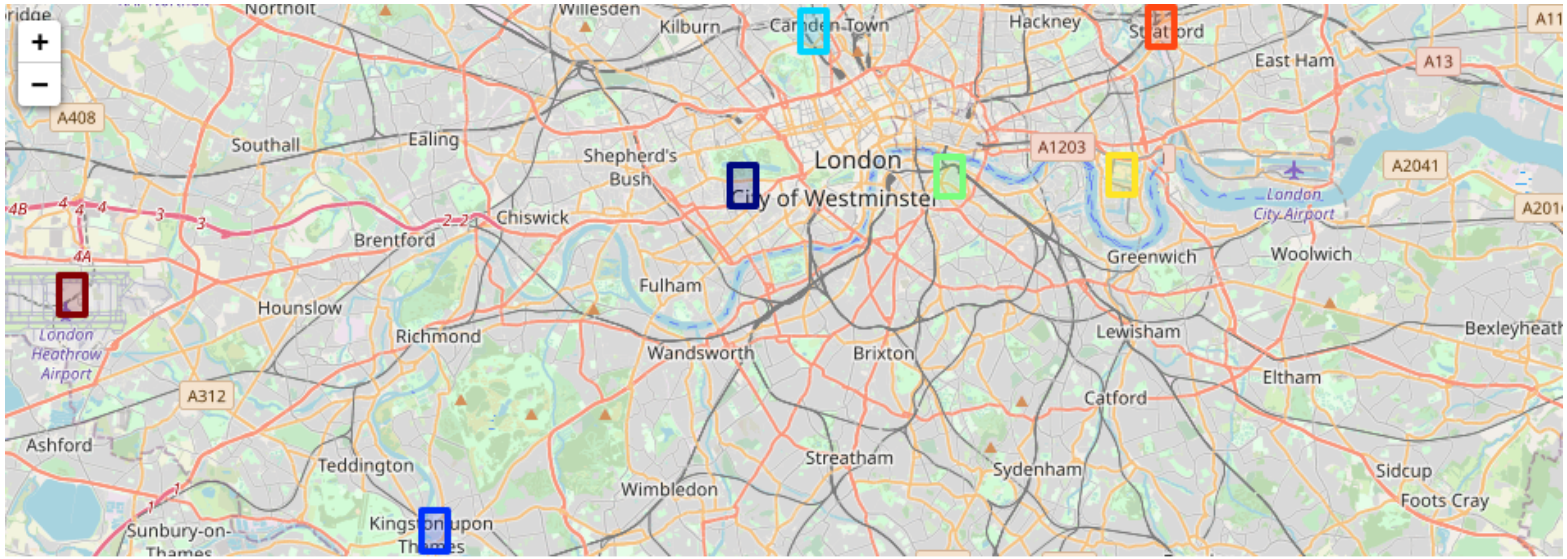


Travel & Transport

Results

- Seven 1 km by 1 km areas of high venue density:





Discussion

Segmenting the Real State Agencies clients according to preferences relative to their way of life, we can make the following recommendatios:

- Clients who give a high value to **Culture and Education**: *area 0, located in Kensington and Chelsea.*
- Clients who wish for a **comfortable life**, with plenty of **opportunities for a good meal**, and **abundance of services** in a **highly residential region**: *area 1, located in Kingston upon Thames*.
- Clients with a **passion for the night life**: *areas 2 (in Camden) and 3 (in Southwark).*
- Clients who prioritize their **careers**: *area 2 (in Camden) and 4 (in Tower Hamlets)*.
- Clients who value access to **Outdoor & Recreation** venues: *area 4 (in Tower Hamlets)*

Conclusion

- We concluded successfully the Capstone, in which we formulated a problem to be tackled with the help of data science concepts, tools and techniques. The project leveraged location data to help solve the problem, as required by the capstone directives. Data from Foursquare, Bing, Postcode.io, and Wikipedia were employed.