Capstone Project – Luxury bakery in London

Applied Data Science Capstone by IBM/Coursera

1. Introduction

1.1. Background

Today an ordinary food becomes a part of art. And bakery/pasties is not an exception. We'd like to offer a luxury bakery for sophisticated London public, who will appreciate this.

1.2. Problem

In this project we will try to find an optimal location for our luxury bakery. Specifically, this report will be targeted to stakeholders interested in opening a luxury bakery in London, UK.

Since there are lots of venues in London we will try to detect locations that are not already crowded with bakeries. We are also particularly interested in areas with no popular bakeries. We would also prefer posh locations for wealthy people.

We will use our data science powers to generate a few most promising neighborhoods based on this criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

2. Data

2.1. Data sources

Based on definition of our problem, factors that will influence our decission are:

- Average monthly rental costs in London. The more such costs, the more wealthy people live in that boroughs. Therefore, it'll be considered the good place for opening a luxury bakery.
- The number of bakeries in top10 places in the chosen boroughs.

We decided to use regularly spaced grid of locations, centered around city center, to define our neighborhoods.

Following data sources will be needed to extract/generate the required information:

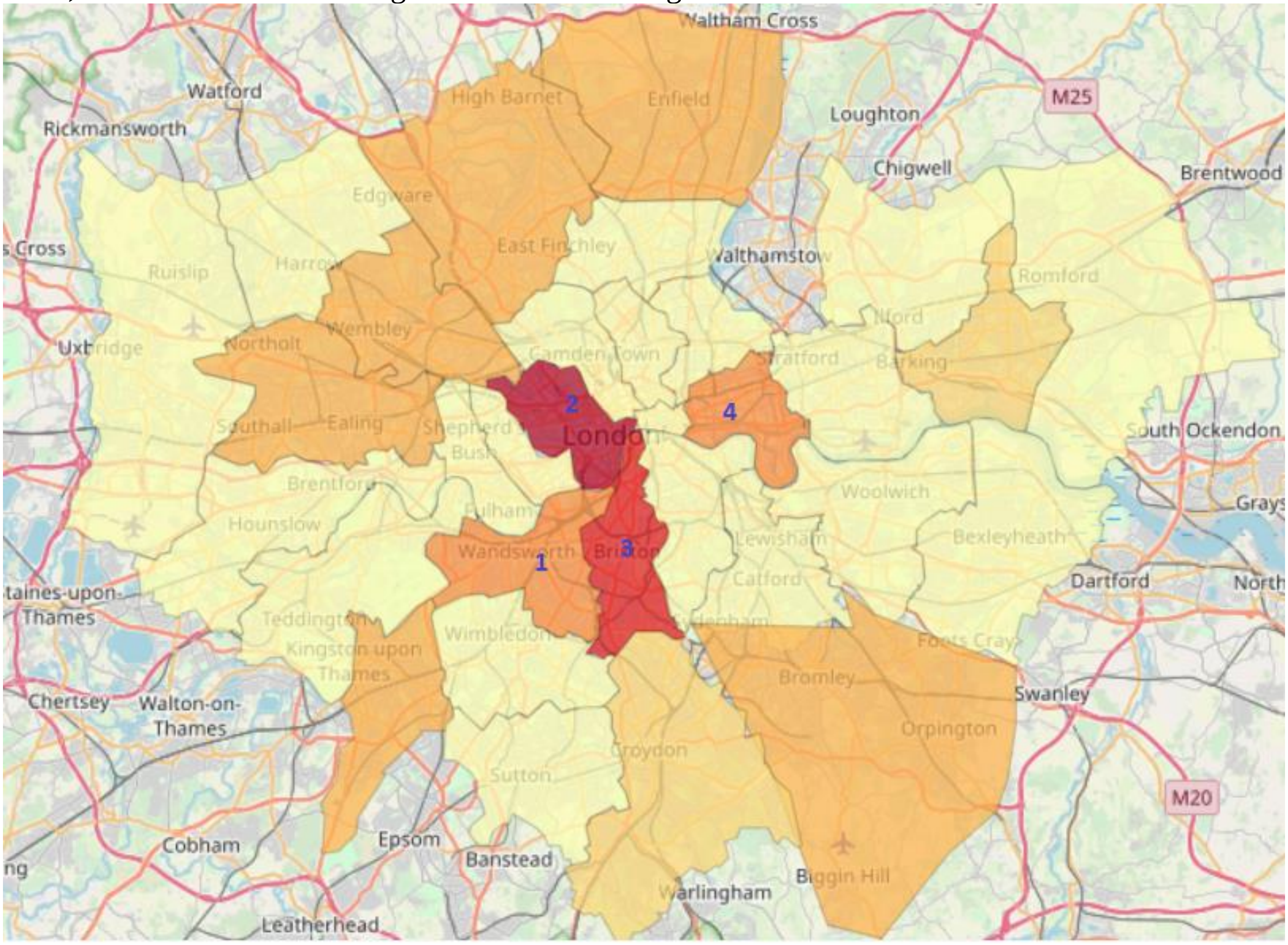
- Average monthly rental costs in Greater London as of June 2019, by borough (in GPB) https://www.statista.com/statistics/752279/average-rental-costs-in-greater-london-boroughs/

- number of top10 places and their type and location in every neighborhood will be obtained using Foursquare API
- Greater London Area postal codes https://en.wikipedia.org/wiki/List_of_areas_of_London . The BeautifulSoup package will be used to scrap the needed data from Wikipedia.

2.2. Data cleaning

We downloaded average monthly rental costs from statista.com. As it's seen the several boroughs were presented in one line what can be a real problem for future execution. That's why we use formula *.split().stuck().unstuck()* to transform the line into several lines.

Then, we chose 4 cental boroughs for further investigation.



For that investigation we need coordinates of these boroughs. In this project, London will be used as synonymous to the “Greater London Area”. Within the Greater London Area, there are areas that are within the London Area Postcode. The focus of this project will be the neighborhoods are that are within the London Post Code area. The London Area consists of 32 Boroughs and the “City of London”. Our data will be from the link — Greater London Area https://en.wikipedia.org/wiki/List_of_areas_of_London . The BeautifulSoup package is used to scrap the needed data from Wikipedia. In order to get the needed data we:

- chose only 4 certain boroughs,
- deleted duplicated postal codes
- rename 'City' to 'City of London'

In obtaining the location data of the locations, the Geocoder package is used with the arcgis_geocoder to obtain the latitude and longitude of the needed locations. These will help to create a new dataframe that will be used subsequently for top4 areas.

3. Methodology

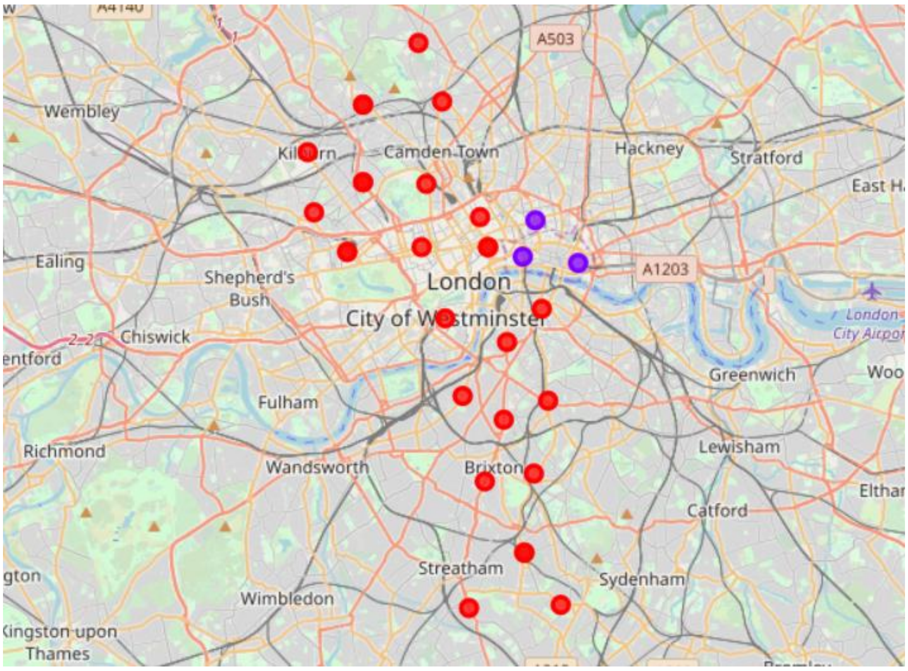
In this project we will direct our efforts on detecting areas of London that have low top bakeries. We have limited our analysis to area ~1km around neighborhoods' centers.

In the first step we have collected the required data: location and type (category) of every venue within 1km from neighborhoods centers.

In the second and final step we will focus on most promising areas and within those create clusters of locations that meet some basic requirements established in discussion with stakeholders. We will present map of all such locations but also create clusters (using k-means clustering) of those locations to identify general zones / neighborhoods, which should be a starting point for final 'street level' exploration and search for optimal venue location by stakeholders.

4. Analysis

Let us now cluster those locations to create centers of zones containing good locations. Those zones, their centers and addresses will be the final result of our analysis.

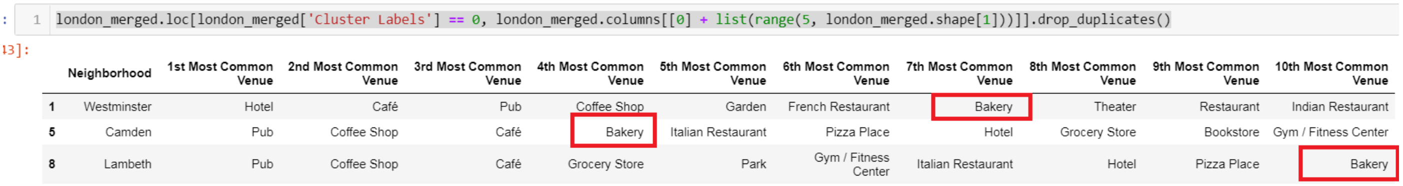


5. Results

Our analysis shows that there are 4 certain neighborhoods that are best suited for offering luxury bakery segment. It was based on the fact that the rent price is positively correlated with the wealth of people leaving in that neighborhoods.

After directing our attention to this more narrow area of interest we first explored them on top10 most popular venues.

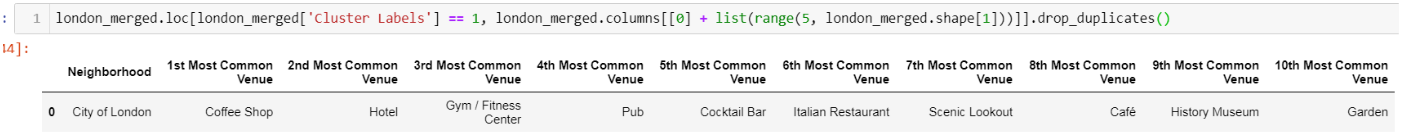
Those location candidates were then clustered to create zones of interest. Addresses of centers of those zones were also generated using reverse geocoding to be used as markers/starting points for more detailed local analysis based on other factors.



```
london_merged.loc[london_merged['Cluster Labels'] == 0, london_merged.columns[[0] + list(range(5, london_merged.shape[1]))]].drop_duplicates()
```

13]:

	Neighborhood	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
1	Westminster	Hotel	Café	Pub	Coffee Shop	Garden	French Restaurant	Bakery	Theater	Restaurant	Indian Restaurant
5	Camden	Pub	Coffee Shop	Café	Bakery	Italian Restaurant	Pizza Place	Hotel	Grocery Store	Bookstore	Gym / Fitness Center
8	Lambeth	Pub	Coffee Shop	Café	Grocery Store	Park	Gym / Fitness Center	Italian Restaurant	Hotel	Pizza Place	Bakery



```
london_merged.loc[london_merged['Cluster Labels'] == 1, london_merged.columns[[0] + list(range(5, london_merged.shape[1]))]].drop_duplicates()
```

14]:

	Neighborhood	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
0	City of London	Coffee Shop	Hotel	Gym / Fitness Center	Pub	Cocktail Bar	Italian Restaurant	Scenic Lookout	Café	History Museum	Garden

After analysis of those 2 clusters we came to conclusion that cluster#1 is less attractive for stakeholders than cluster #2. It's explained by the fact that all of the neighborhoods in Cluster#1 has bakery in their top10 most common venues whereas Cluster#2 doesn't have at all. It means that the City of London should be the starting point for more detailed analysis which could eventually result in location which has not only no nearby competition.

6. Conclusion

Purpose of this project was to identify London neigborhood close to center with low number of bakeries in top venues in order to aid stakeholders in narrowing down the search for optimal location for a new luxury bakery. We chose top4 areas based on the highest average month rent, assuming that rent price is positively correlated with the wealth of people leaving in that neighborhoods.

Clustering of those neighborhoods was then performed in order to create major zones of interest (containing greatest number of potential locations) and addresses of those zone centers were created to be used as starting points for final exploration by stakeholders.

Final decission on optimal restaurant location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.