BATTLE OF THE NEIGHBORHOODS

OPENING RESTAURANT IN THE GREATER LONDON AREA

Applied Data Science Capstone Project

Author: Ksenia Tabakova

Abstract: Greater London area wards (neighborhoods) were studied in order to identify the best areas to open restaurants that serve either American, Mexican or Italian cuisine. Additionally, price range recommendation for potential new restaurants was developed. Three type of wards were identified: (1) relatively low medium income, relatively few restaurants per 1000 people – 58% of wards; (2) more restaurants per 1000 people, varying income – 13% of wards; (3) higher medium income, relatively few restaurants per 1000 people – 29% of wards. For every ward type following recommendations were developed: (1) – Mexican restaurant, lower price tag on menu; (2) – Mexican and American restaurants, lower and higher price tag on menu possible; (3) Mexican and American restaurants, higher price tag. Top 10 wards based on the predicted population growth were identified for each type of ward.

Table of Contents

1. Introduction

A prominent restaurant entrepreneur in UK that runs restaurants serving Italian, American and Mexican dishes thinks of opening more restaurants in Greater London area to increase business coverage. London is known for its vivid cuisine scene and there is abundance of the restaurants and food joints satisfying all tastes. The question is where to open and which restaurant to open? There are several parameters that can help assessing future of the newly opened restaurant:

1. **Abundance of similar restaurants** in the same area. For instance, it would be unwise to open Italians restaurant in the area where there are already many Italian restaurants open, particularly if they have high rating. It will be hard to compete and may cause monetary losses. On the other hand, bringing new tastes to the area is potentially a recipe for success.

2. Another factor for decision is **population/housing density** - suburban districts oftentimes are dominated by the houses for 1 family and can be characterized by the low population density. Such districts are not able to support many small businesses - there will be not enough clients to achieve profit. Open areas and parks in the vicinity of the neighborhood will impact population density as well. European suburban areas are characterized by higher population density than, for instance, US, but relative population density nonetheless plays role in formation on neighborhood dynamics and citizen flow.

3. **Income** of citizens can also determine whether restaurant will be successful or not. Higher income in the neighborhood would be beneficial for a new restaurant.

4. **Population growth** - if population is expected to grow, restaurant will be able to serve more customers. Area where population is expected to decrease would be a worse investment with higher risks.

Entrepreneur turns to me, an aspiring data scientist, and asks to provide data-based answers for the **three main questions**:

- Which Greater London areas are the best to open restaurant(s)?

- Which cuisine(s) (American, Mexican or Italian) are the best option to open in the identified areas?

- What price range for newly opened restaurant should be considered (low or high price tag)?

2. Data sources and description

2.1 Data sources

Greater London Area that I studied has 32 local government districts called **borough** plus special district of City of London. Each borough is divided by **wards** which serve as primary electoral geographical unit. I used wards as a proxy for neighborhoods. I excluded City of London from the analysis and concentrated on all other areas of inner and outer London.

For this project I utilized following datasets:

1. Greater London Area (GLA) Land Area and Population Density, Ward and Borough. This open dataset is based on the census from 2011, as well as projections of population till year 2050. Variables used:

    - Total population for year 2021 (note that since last census took year in 2011, I need to use projection data)
    - Population density for year 2021 (calculated for inland area, does not include water bodies).
    - Projected population by 2026 to compute population growth which serves as good indicator to select prospective wards: if population will grow, more potential customers restaurant will have.

2. GIS data on boundaries of wards in 2011. I used shapefiles containing data on geometry(polygons) of wards as they were in 2011 to find centroids of each ward that I need to call Foursquare API.

3. **Foursquare API** to fetch information about venues in each of the wards. I used coordinates obtained from the dataset #2. I extracted information about type of the venue, only calling venues that serve food (cafes, fastfood joints, restaurants etc).

4. Income data for wards. I choose to use median income over mean as better metric to judge overall ward income as it is shows most common income in a ward.

5. Open space data for each ward - I used this metric it to calculate adjusted population density that does not take into account vast uninhabited place, such as parks, stadiums etc.

Information on authors of datasets and licenses can be found in References section.

2.2 Data samples

All datasets used in this study have unified system for coding wards. Since 2011 many wards experienced changes (recoding, changed area, merge with other wards *etc*), and codes from 2011 don't always match all current codes. For this study I ensured that I use data with the same coding. This made data manipulation easy and practical. Mostly I just had to filter necessary data columns, do type conversion, drop missing values.

Let's have look into samples of all datasets (after cleaning and selecting appropriate columns).

1. Population data

	Code	Borough	Ward_Name	Year	Population	km2	Population_density
6240	E05000026	Barking and Dagenham	Abbey	2021	16938	1.279	13243.158720
6241	E05000027	Barking and Dagenham	Alibon	2021	11323	1.361	8319.617928
6242	E05000028	Barking and Dagenham	Becontree	2021	14891	1.284	11597.352020
6243	E05000029	Barking and Dagenham	Chadwell Heath	2021	11297	3.380	3342.307692
6244	E05000030	Barking and Dagenham	Eastbrook	2021	11032	3.454	3193.977997

Figure 1. *Sample of UK census 2011 data on wards' population.*

Code column (Figure 1) was used as main identificator. Code is provided for every dataset and can be used to merge dataframes. Note: area of each ward does not include water bodies.

The same dataset contains population prediction up to year 2050. I selected year 2026 to calculate expected population growth within next 5 years for every ward which was used to select the best ward candidates.

2. Shapefiles of wards (as of 2011).

	Code	geometry
0	E05000405	POLYGON ((516401.600 160201.800, 516407.300 16...
1	E05000414	POLYGON ((517829.600 165447.100, 517837.100 16...
2	E05000401	POLYGON ((518107.500 167303.400, 518114.300 16...
3	E05000400	POLYGON ((520480.000 166909.800, 520490.700 16...
4	E05000402	POLYGON ((522071.000 168144.900, 522063.900 16...

Figure 2*. Sample of shapefile data (left) and example of polygon geometry data (right).*

Column geometry contains information about polygon shape (Figure 2). I used **geopandas** library to compute centroid for each polygon/ward. I used centroid coordinates for API calls. However, geometry as well as centroids are provided as easting and northing in meters

(meters from 0 longitude, 0 latitude, Figure 3). I needed to convert meters to degrees. I used **pyproj** library to convert between meters and degrees.

	Code	geometry	Centroid
0	E05000405	POLYGON ((516401.600 160201.800, 516407.300 16...	POINT (517652.343 162339.161)
1	E05000414	POLYGON ((517829.600 165447.100, 517837.100 16...	POINT (519124.935 165300.017)
2	E05000401	POLYGON ((518107.500 167303.400, 518114.300 16...	POINT (519108.407 167344.325)
3	E05000400	POLYGON ((520480.000 166909.800, 520490.700 16...	POINT (520118.140 166393.329)
4	E05000402	POLYGON ((522071.000 168144.900, 522063.900 16...	POINT (521204.946 168516.788)

***Figure 3**. Sample of centroid data for each ward, in meters of easting and northing. before conversion to degrees of latitude and longitude.*

3. Median income per ward (pounds).

```
income.head()
```

	Code	Income
0	E09000001	63620
1	E05000026	33920
2	E05000027	32470
3	E05000028	33000
4	E05000029	33920

I used latest available income data from years 2012/2013. It was the latest income estimation for the wards as they were in 2011. Clearly, income distribution between wards might have changed since 2013, but it should serve as a good proxy for current income distribution.

***Figure 4.** Sample of median income data.*

4. Open space data (in %).

```
space.head()
```

	Code	open_space
0	E05000026	21.908601
1	E05000027	20.621849
2	E05000028	1.885448
3	E05000029	55.974507
4	E05000030	50.434179

I used information on open area percent per ward (parks, golf fields, stadiums) to calculate adjusted population density (Figure 5). I hypothesize that if a given ward has lots of open space, population density number can be too low and skew results. Adjusted population density is simply:

$$Density_{adjusted} = \frac{Population}{area * open\_space} * 100$$

***Figure 5.** Sample of open space (parks, stadiums, golf courses etc) in % for wards.*

2.3 Foursquare API calls

2.3.1 Calls configurations

Foursquare allows to make targeted calls, requesting data only for certain types of venues. I made **'explore'** calls for food venues (**categoryIDs** endpoint 4d4b7105d754a06374d81259) within 750 m radius of every ward's centroid. For some wards I had to make second call to fetch second page of venues (if a ward has more than 100 venues in the radius).

2.3.2 Foursquare restaurant data

The client has restaurants that serve dishes from either American, Mexican or Italian cuisines. There were 137 unique categories of food venues found for wards of Greater London. I classified all venues to be one of the following types:

- **American:** ['Fried Chicken Joint', 'Diner', 'American Restaurant', 'Wings Joint', 'Bagel Shop', 'Burger Joint', Australian Restaurant', 'Southern / Soul Food Restaurant', 'Cajun / Creole Restaurant', 'BBQ Joint']. Note: I included Australian restaurants in this category as they normally serve similar food, such as burgers, lobster, steaks.
- **Mexican:** ['Burrito Place', 'Mexican Restaurant', 'Taco Place'].
- **Italian:** ['Pizza Place', 'Italian Restaurant', 'Veneto Restaurant'].
- **Other** – everything which is not one of the above.

I have performed one hot coding for the categorical variable **'Type'** and have calculated fraction of American, Italian and Mexican restaurants of total number of restaurants/food venues in every ward.

2.4 Method selection

The questions posed in the project demand to perform unsupervised clustering: the data is unlabeled, we don't know in advance which wards are similar. I have selected k-means clustering in order to find best ward candidates and to give detailed recommendations to the client.

2.4.1 Features selection

In order to select feature set for clustering, I have explored whether hypothesized features exhibit any connections and have potential for clustering (Figure 6). There are several important observations I can make about the data:

1. We can see that there doesn't exist obvious correlation between selected parameters. For instance, higher population density does not necessarily go hand in hand with more restaurants per 1000 citizens. Neither income directly impacts number of restaurants per 1000 citizens. In some cases lower income wards have high number of restaurants/1000 people – people inhabiting such wards go to eat out regardless of disposable income.
2. Income and population display some connection - the higher median income of the ward, the fewer people inhabit it: we attribute it to be suburban districts where housing is scarcer because people own/rent private houses, whilst private house is more expensive property to own/rent.
3. Adding adjusted population density did not yield desired/expected result - there is still not clear connection with restaurant number per 1000 people (neither with absolute number of restaurants, not shown here).

4. From the business point of view, income and restaurants/1000 people are important parameters to make decision over. Adding any extra parameters which do not affect those would be unnecessary and wasteful in terms of computational and analysis time.

Most importantly we can see that population and restaurants number (per 1000 people as well as total number restaurants, not shown here) do not correlate. Data look like a cloud. There is some indication of bimodality of population distribution. In fact, I have tried various combinations and filter over population data, but did not gain any new insights.

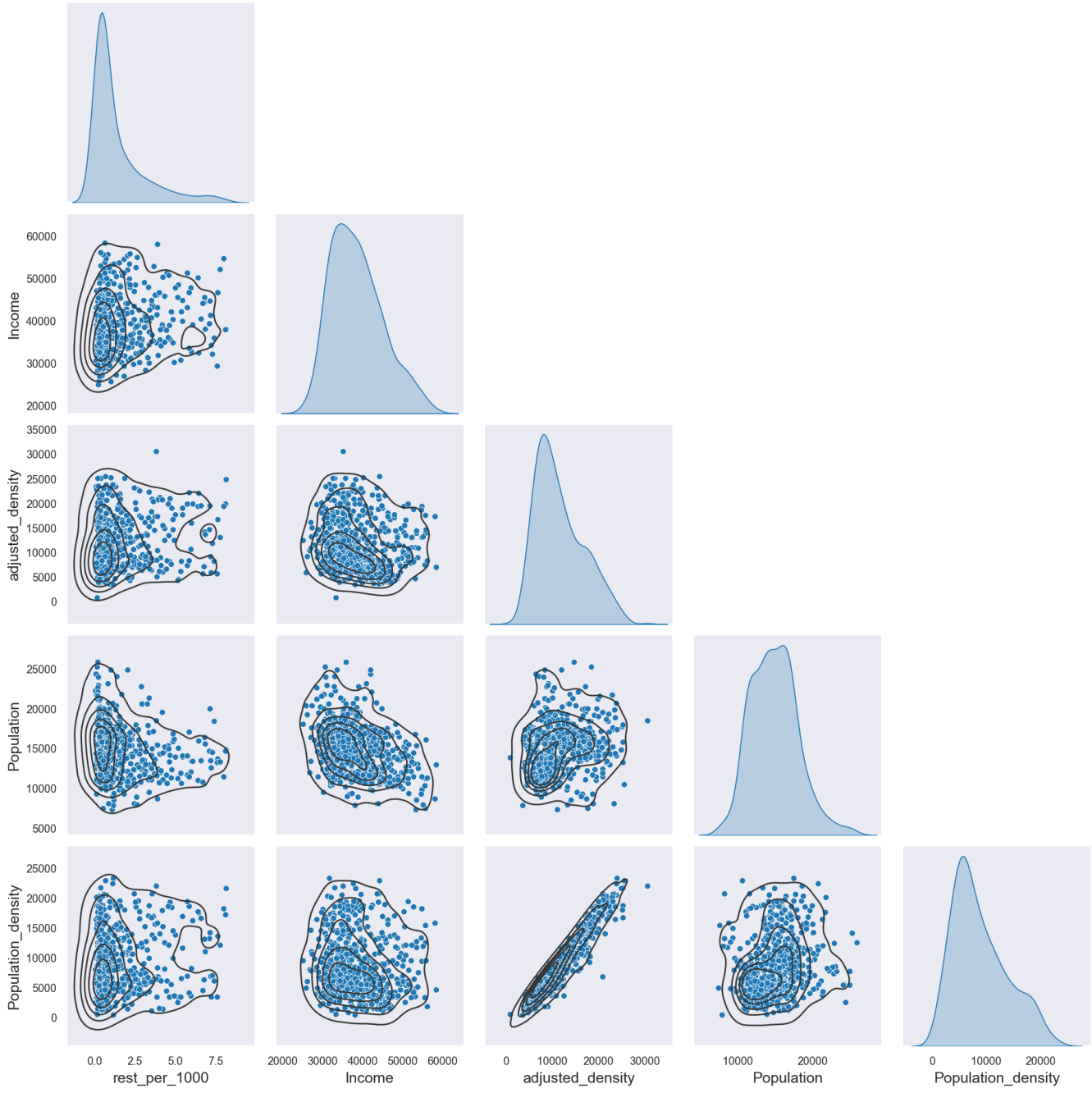

***Figure 6**. Exploratory data visualization of potential features for k-means clustering. Lines on scatter plots show kernel density estimates (KDE)*

To conclude, I selected two parameters for kmeans clustering - **Income** and **# of restaurants per 1000 people** - to aid decision making. Some wards clearly have bustling restaurant scene and we hope we can discriminate between bustling/non-bustling and then look further on the distribution of types of restaurants.

2.4.2 k-value selection

I used elbow method (on Euclidian distance) to decide the best number of clusters (Figure 7).

Figure 7. *Elbow curve.*

Unsurprisingly, elbow curve is rather smooth – the data did not visually show any clear signs of being clustered. There are two candidates for optimal k-value – 3 and 5. I have performed clustering for both of these k-values, and higher k-value did not improve results or made recommendation-making easier. I select k=3 for the further analysis.

3. Exploratory data analysis

3.1 k-means clustering results

Results of clustering are displayed in Figure 8.

Clusters can be described as follows:

- **Cluster 0**: generally lower median income, relatively fewer restaurants per 1000 people. 58% of wards belong to this cluster.
- **Cluster 1**: varying income, more restaurants per 1000 people. These are wards with lively dining life. 13% of wards belong to this cluster.

- **Cluster 2**: generally higher income, fewer restaurants per 1000 people. 29% of wards make up this cluster.

It is important to note that there is overlap between clusters: clusters 0 and 1 overlap at the higher income portion of data, while clusters 0 and 2 overlap at the restaurant number per 1000 people. It is definitely expected and not optimal, but this is real-life data. However, there is enough separation for income distributions, as well as clear difference between wards with lively and not-lively restaurant life.

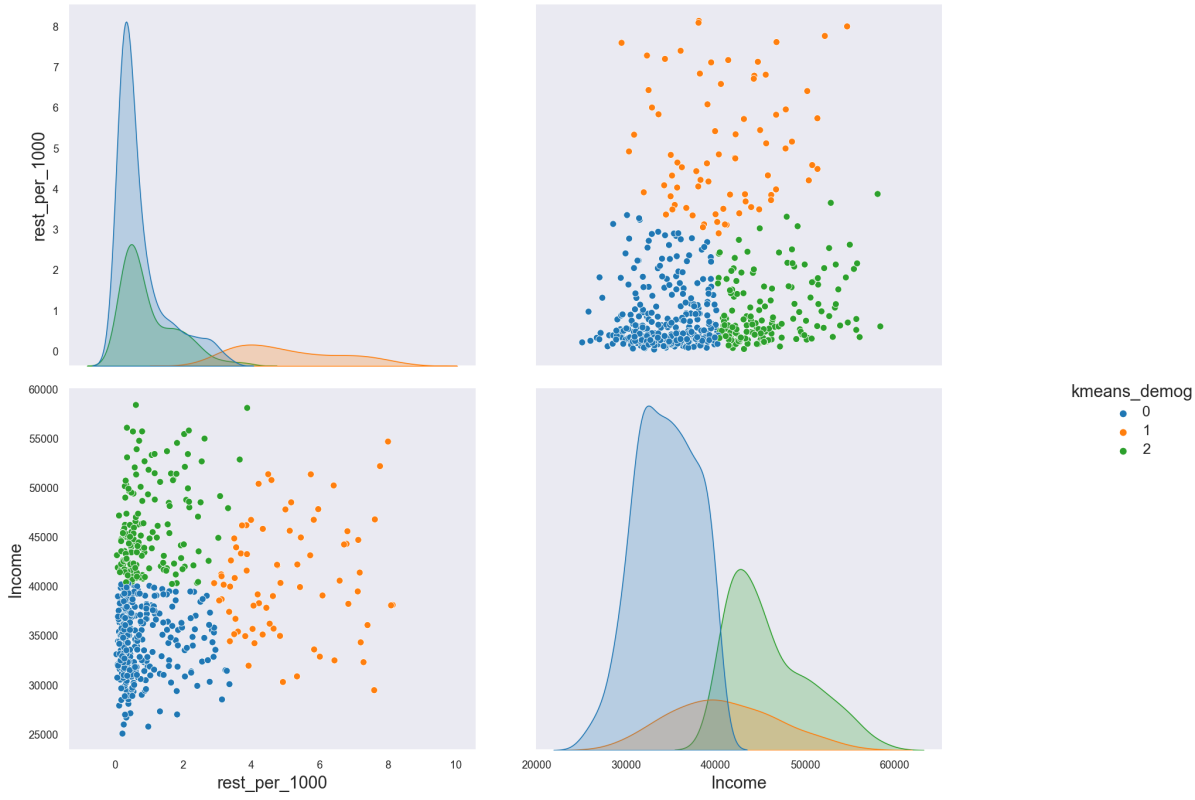

***Figure 8**. Visualization of clustering results.*

3.2 Candidate wards - reasoning

Let us discuss type of wards that would be in theory the best candidate to open restaurants.

- Wards with higher income and less lively restaurant life: people have money to spend but restaurant life in the neighborhood is not saturated. **Cluster 2.**
- Wards with lively restaurant life - people are going out to eat regardless of the income. Habits of people show that there is possibility compete for clients. Type of restaurant will be decisive factor. **Cluster 1**.
- Wards with not lively restaurant life and smaller income - there is a potential for success if type of restaurant is considered carefully. **Cluster 0.**

3.3 Distribution of American, Mexican and Italian restaurants in clusters.

To answer posed questions, I have studied cluster-wise distribution of restaurants serving American, Mexican and Italian cuisines. I have calculated mean fraction of above-mentioned restaurants and have plotted it for each cluster (Figure 9).

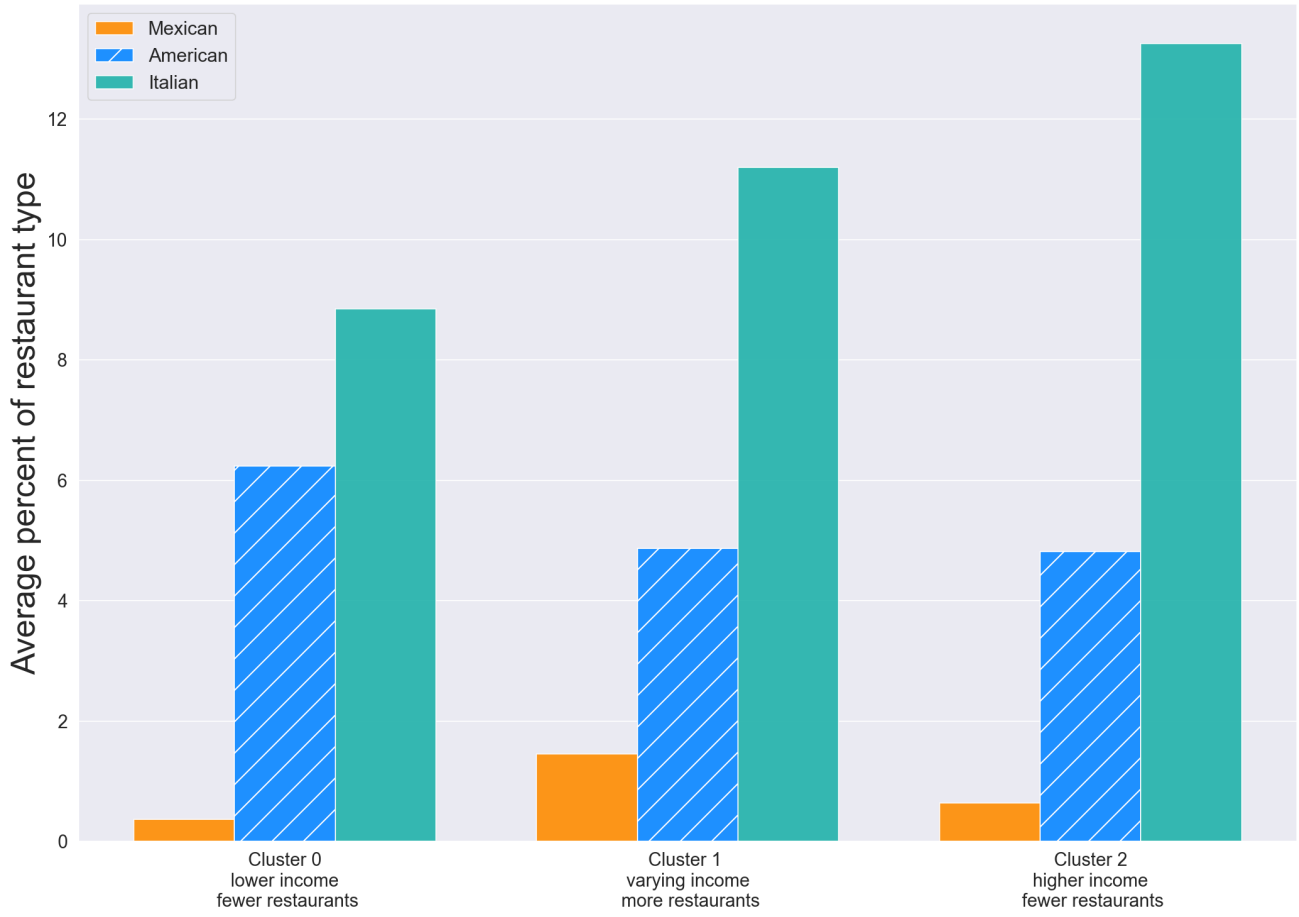

***Figure 9**. Distribution of American, Mexican and Italian cuisine restaurants in clusters.*

We can start making first recommendations about which restaurants client can consider opening in wards of a given cluster.

- Mexican restaurants are in minority overall, regardless of the cluster. It means that our client can seriously consider opening Mexican restaurants, particularly in wards of Clusters 0 and 2.
- Italian restaurants are most prevalent across wards, taking up from 9 to 13 % of total number, depending on the ward. Competition is high in this segment. We will not recommend opening more restaurants in either of clusters in this analysis.
- American cuisine represents 4-6% of all food venues across wards. There is a possibility to add more American restaurants in Clusters 1 and 2.

Let me summarize recommendations by clusters and add menu price tag option.

- **Cluster 0** (lower income, fewer restaurants per 1000) - Mexican; lower price tag.
- **Cluster 1** (lively restaurant life regardless of income) - Mexican, American; both lower and higher price tag possible.
- **Cluster 2** (high income, fewer restaurants per 1000) - Mexican, American, higher price tag possible.

On the map clusters are distributed in the following way:

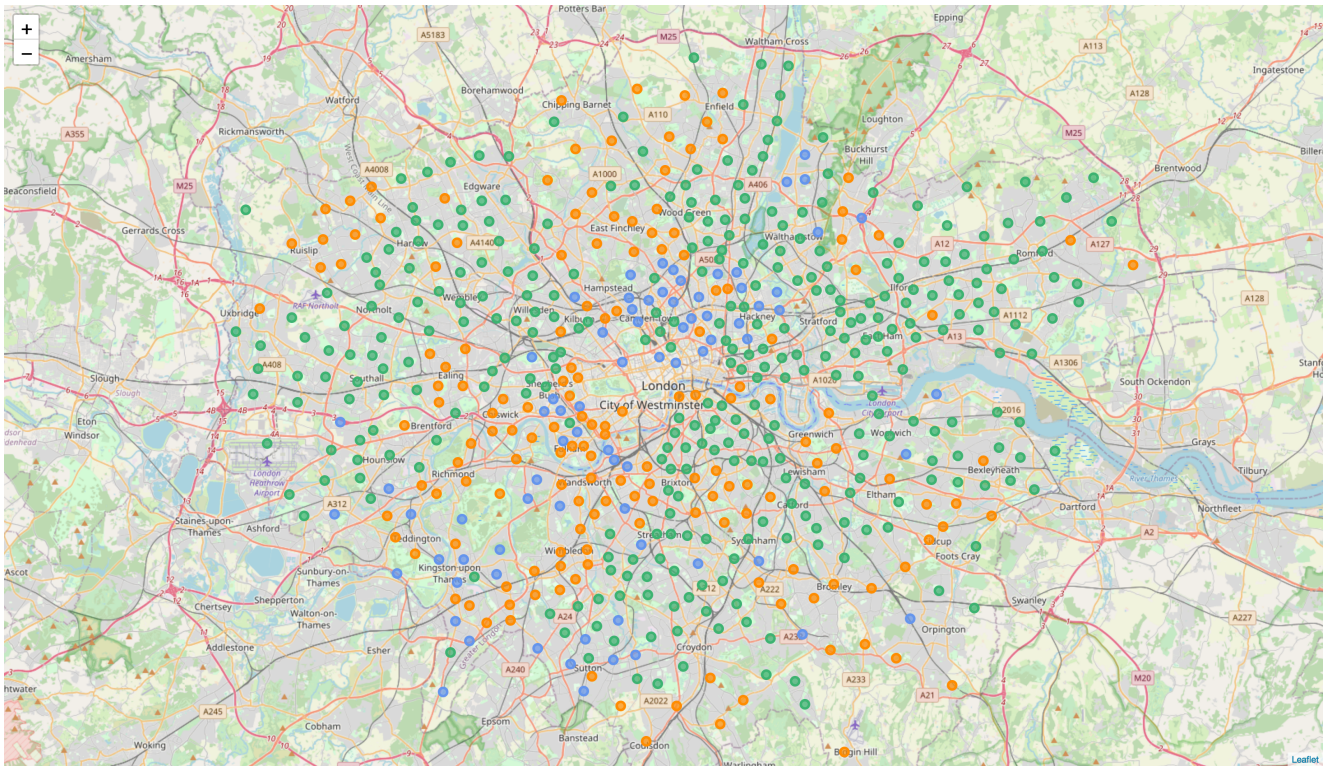

Figure 10. *Geographical distribution of clusters.*

3.4 Population growth

Within every cluster I have sorted wards based on the highest expected population growth: the higher the growth, the better survival rate of new venues we can expect. Geographic distribution of top 10 wards from each cluster are shown on Figure 11.

- **Cluster 0** – lower income, fewer restaurants (Table 1). Top 10 wards all show favorable expected population growth, with #1 going as high as 41%. Lowest expected growth is 27% which is also high. #2 and #4 are from the same borough, Ealing, making this borough very good candidate overall.
- **Cluster 1** – varying income, more restaurants (Table 2). Expected population growth of top 10 wards in this cluster varying from 8 to 72%, and top one ward exceed second best wards by 50%. Top two wards are from the same borough. Perhaps this area

expected to be under development or redevelopment, with new residential housing being built. These two wards are very strong candidates.

- **Cluster 3** – higher income, fewer restaurants (Table 3). Range of expected population growth is narrower than in Cluster 0 and 1 – varying from 10 to 34 %. #3 and #7 are from the same borough Merton, suggesting that it is a good candidate.

	Ward_Name	Borough	recommendation	Growth
0	Tokyngton	Brent	Mexican; lower price tag	41.209948
1	Southall Broadway	Ealing	Mexican; lower price tag	36.752350
2	Upper Edmonton	Enfield	Mexican; lower price tag	34.864208
3	East Acton	Ealing	Mexican; lower price tag	33.780928
4	River	Barking and Dagenham	Mexican; lower price tag	31.392330
5	Royal Docks	Newham	Mexican; lower price tag	30.947766
6	Woolwich Riverside	Greenwich	Mexican; lower price tag	29.156801
7	Shepherd's Bush Green	Hammersmith and Fulham	Mexican; lower price tag	27.902053
8	Livesey	Southwark	Mexican; lower price tag	27.450753
9	South Hornchurch	Havering	Mexican; lower price tag	27.143833

***Table 1**. Top 10 recommended wards for Cluster 0.*

	Ward_Name	Borough	recommendation	Growth
0	College Park and Old Oak	Hammersmith and Fulham	Mexican, American; lower-higher price tag	74.080560
1	Hammersmith Broadway	Hammersmith and Fulham	Mexican, American; lower-higher price tag	22.425598
2	Grove	Kingston upon Thames	Mexican, American; lower-higher price tag	14.137752
3	Wick	Hackney	Mexican, American; lower-higher price tag	11.865915
4	Thamesmead Moorings	Greenwich	Mexican, American; lower-higher price tag	11.692020
5	Tolworth and Hook Rise	Kingston upon Thames	Mexican, American; lower-higher price tag	11.020060
6	Abingdon	Kensington and Chelsea	Mexican, American; lower-higher price tag	10.816220
7	Petts Wood and Knoll	Bromley	Mexican, American; lower-higher price tag	8.571807
8	Hanworth Park	Hounslow	Mexican, American; lower-higher price tag	8.273296
9	St George's	Islington	Mexican, American; lower-higher price tag	8.026685

***Table 2**. Top 10 recommended wards for Cluster 1.*

	Ward_Name	Borough	recommendation	Growth
0	Peninsula	Greenwich	Mexican, American; higher price tag	33.959538
1	Golders Green	Barnet	Mexican, American; higher price tag	26.228589
2	Uxbridge North	Hillingdon	Mexican, American; higher price tag	17.434613
3	Merton Park	Merton	Mexican, American; higher price tag	16.453983
4	Bromley Town	Bromley	Mexican, American; higher price tag	15.143942
5	Mill Hill	Barnet	Mexican, American; higher price tag	13.643449
6	Fairfield	Wandsworth	Mexican, American; higher price tag	12.936550
7	Wimbledon Park	Merton	Mexican, American; higher price tag	10.707833
8	Coulsdon West	Croydon	Mexican, American; higher price tag	10.362507
9	Bishop's	Lambeth	Mexican, American; higher price tag	9.698699

***Table 3**. Top 10 recommended wards for Cluster 2.*

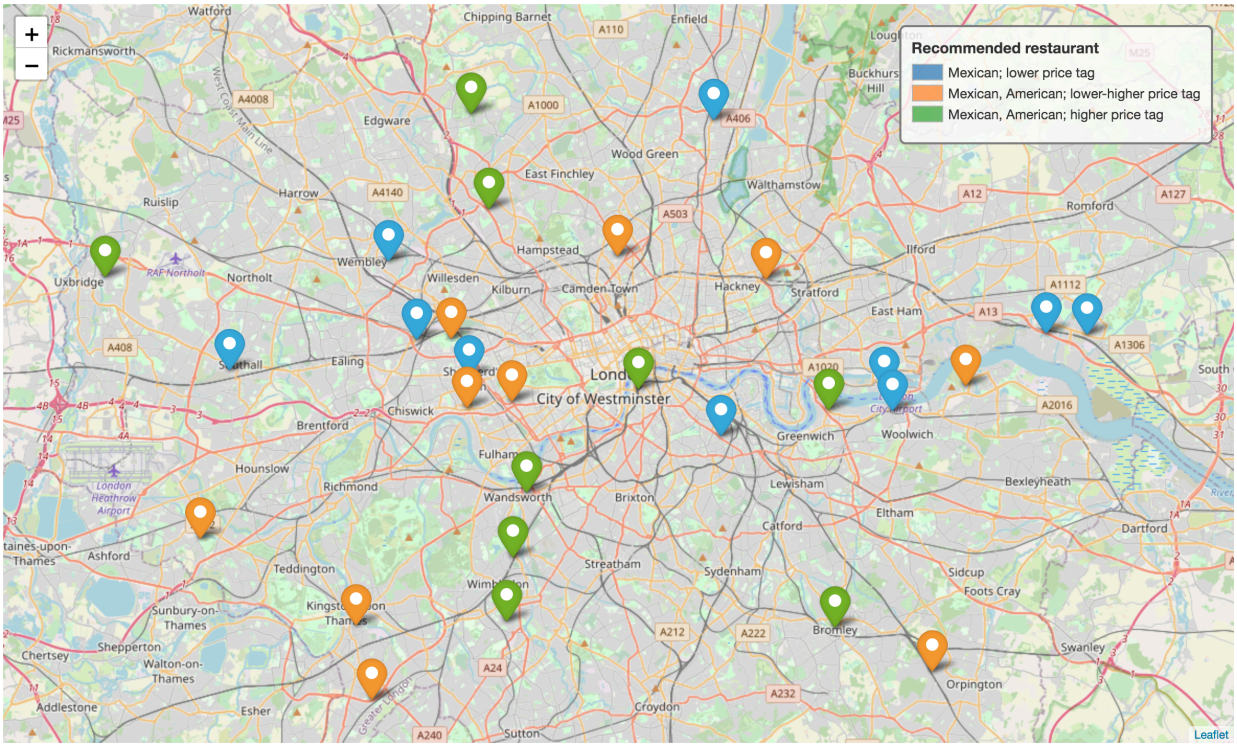

***Figure 11.** Geographic distribution of top 10 wards of each cluster.*

4. Discussion

K-means clustering on median income and number of restaurants per 1000 people have produced three clusters of wards: (1) relatively lower income, fewer restaurants; (2) varying income, more restaurants; (3) higher income, fewer restaurants. Despite some overlap between cluster characteristic, these clusters have different enough properties to be further utilized in the recommendation-making.

Looking into distribution of American, Mexican and Italian cuisine revealed that Mexican restaurants are in minority, therefore making it good candidate for a new restaurant. On the other hand, the fact that Mexican cuisine is not as common as the rest might point that people prefer other types of restaurants, other cuisines. But from variety point of view Mexican restaurants can bring the most of it to a ward.

American cuisine restaurants (including diners, burger joints) take from 4 to 6% of the restaurants in every cluster. Potentially, American restaurant can gain success in either of the clusters, but narrowing down focus to clusters 1 and 2 is the most sensible.

Italian restaurants are the most popular type of restaurants from considered: from 9% in Cluster 0 to 13% in Cluster 2. The competition in this sector is highest. I do not recommend opening Italian restaurants to the client.

Median income of cluster will be the basis of recommendation for the price range:

Cluster 0: lower priced menu.

Cluster 1: both lower and higher priced menu is possible.

Cluster 2: higher price menu.

As it happens often with the real-life data, this project was not as straightforward as expected. I hypothesized that population density has an effect on the total number of restaurants or number of restaurants per 1000 people. Even though such dependency would be logical (the higher the density of populations/customers, the more small businesses including restaurants it can support), there seems to exist other factor(s) that impacts number of restaurants. It might be population distribution (i.e. age distribution), or number of non-residential buildings that can host restaurant. Either way, our data could not help identifying this parameter.

On the positive side, clustering based on the median income and number of restaurants per 1000 people yielded satisfactory results. It was possible to developed data-based recommendations about location, type and menu price range for a restaurant.

5. Conclusions

In this study I analyzed wards (proxy for neighborhoods) in the Greater London area to determine which areas are the best to open either American, Mexican or Italian cuisine restaurant. Additionally, I determined which price range for menu would be the most suitable. I have performed k-means clustering to classify wards. I have developed recommendation on

what type of restaurant (cuisine and price range) can be open at which wards. Using predicted population growth during next 5 years I selected top 10 best candidates for each cluster (30 best wards in total).

6. References

1. **Land Area and Population Density, Ward and Borough**.
   https://data.london.gov.uk/dataset/land-area-and-population-density-ward-and-borough
   Author: Greater London Authority.
   Licence: UK Open Government Licence (OGL v3)
2. **Statistical GIS Boundary Files for London.**
   https://data.london.gov.uk/dataset/statistical-gis-boundary-files-london
   Author: Greater London Authority
   Licence: UK Open Government Licence (OGL v2)
3. **Household Income Estimates for Small Areas.**
   https://data.london.gov.uk/dataset/household-income-estimates-small-areas
   Author: Greater London Authority
   Licence: UK Open Government Licence (OGL v2)
4. **Access to Public Open Space and Nature by Ward**
   https://data.london.gov.uk/dataset/access-public-open-space-and-nature-ward
   Author: Greenspace Information for Greater London
   Licence: UK Open Government Licence (OGL v2)