analysing the most suitable London area for opening an english restaurant

Capstone Project - ML

Sebastian Moldovan

Table of Contents

[EXECUTIVE SUMMARY 1](#_Toc36723109)

[BACKGROUND AND PROBLEM 1](#_Toc36723110)

[INTEREST 2](#_Toc36723111)

[DATA ACQUISITION AND DATA CLEANUP 2](#_Toc36723112)

[DATA CLEANING 2](#_Toc36723113)

[CLASSIFICATION 4](#_Toc36723114)

[INTERMEDIARY CONCLUSIONS 9](#_Toc36723115)

[CLUSTER ANALYSIS 9](#_Toc36723116)

[FINAL CONCLUSIONS 15](#_Toc36723117)

[FUTURE INVESTIGATIONS 17](#_Toc36723118)

# Executive summary

Using data related the London areas like population density, median income, employment, happiness level, diversity (population born abroad), business sustainability, restaurants/bars/coffee shops numbers, restaurant type we can split the London areas into several clusters based on similitude. Analyzing each cluster we concluded that two of these clusters are well suited for opening an English restaurant these two includes areas like: City of London, Greenwich, Islington, Kensington and Chelsea, Tower Hamlets. The conclusions were made based on cluster attractiveness (relative higher median income area, age distribution, business success rate, existing number of restaurants and attractiveness of English restaurants).

Further analysis would consist in adding restaurants rankings as a feature of classification and have a deeper dive into the above mentioned areas and select the most suitable one for opening a new restaurant.

# Background and Problem

When it comes to restaurants London is one of the most attractive and competitive place for restaurants. With high population (9 Mil) and large numbers of tourists (30 Mil per year), the capital of England host approximately 18,000 restaurants. Considering the high cost of opening a restaurant (500-1,000K$ the need of having some data driven decision seems imperial. The location of the restaurant will impact the success rate along with the cost allocated. For this we will try to identify the best location to open an English restaurant by trying to segment the London areas and see if there are some similitudes between them. Once they are identified we can start identifying the best cohort and choosing the optimal location where a restaurant will work.

# Interest

Considering the challenging London restaurants landscape it would be very interesting to see if we can analyze the London areas and find the best unexplored areas for opening a new restaurant. The analysis should provide insights on the attractiveness of some areas and should guide food startups on which area are the best place to invest their money in order to maximize their chances of success.

# Data Acquisition and Data Cleanup

The primary data source was taken from Kaggle and it was enhanced with venues information taken via foursquare API. The Kaggle data source contain crucial info like: population density, median income, population, population diversity, housing prices, crime rate, business survival rate, largest migrant population, etc while the foursquare data comes to add the venues around these areas. For simplicity the venues sample is 100 for each area.

## Data cleaning

The Kaggle data cleaning consist of changing the data format and elimination of particular symbols that are making the data unusable. Eliminating columns that are not adding value to our exercise by identifying redundant columns.

Several actions were taken in order to clean the data.

* Redundant features that have a very high correlation factor were eliminated. Correlation factor >0.9
  + households and population have a very high correlation so households will be dropped
  + median income and house price have a high correlation so house price will be dropped
  + life satisfaction and happiness have a high correlation so happiness will be dropped

A picture containing drawing, food

Description automatically generated

Figure 1: Correlation of the features

* In an alternative calculation (results available in FINAL\_TABLE\_2['Cluster\_Reduced2']) of the clusters the items that have a significant correlation with number of restaurants were selected
  + Selected features: 'Job\_Density', 'Median\_Income\_H','Anxiety', 'Active\_Business', 'Bars#', 'Coffee#’, 'Extra#', 'Suvirval\_Rate\_2Years' 'Population'
* In addition to data cleaning for all areas were calculated the following
  + Number of venues in each area (Used in Classification)
    - Number of restaurants
    - Number of bars
    - Number of coffee shops
  + Top 10 restaurants for each area (Used in Classification)
  + Top 10 venues categories for each area (Not Used in Classification)
  + Top 3 nationalities (Used in Classification)

## Classification

The method used for determining the classification of the data set was k-mean classification, due to its simplicity and fitness for this kind of unsupervised approach. The initial number of cluster was following the SRQT(N/2) recommendation but the final cluster number selected was 5.

Multiple variances of classification were calculated and the resulted along with a description was added in Table 1 below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataframe | Scenario | Description | Used | Distribution |
| alternative\_cluster/ alternative\_cluster | 1 | * Inner London * All Venues * No Migrants   Features   * 'Population' * 'Households', * 'Population\_Density' * 'Working\_Population' * 'Youth\_Population', * ‘Elderly\_Population' * 'Born\_Abroad' * 'Employment' * 'Unemployment' * 'Annual\_Pay' * 'Median\_Income\_H' * 'Job\_Density', 'Active\_Business' * ‘Suvirval\_Rate\_2Years' * 'Crime' * 'House\_Price' * 'Life\_Satisfcation' * 'Happiness' * 'Anxiety' * Top10 Venues | Maybe | A picture containing text, map  Description automatically generated |
| A / X\_A | 2 | * Inner London * No Venues * No Migrants   Features   * 'Population' * 'Population\_Density', * 'Working\_Population', * Youth\_Population' * 'Elderly\_Population', * Born\_Abroad', * 'Employment', * Median\_Income\_H', * Job\_Density' * 'Active\_Business', * 'Suvirval\_Rate\_2Years' * 'Life\_Satisfcation' * 'Anxiety’ | NO | A picture containing text, map  Description automatically generated |
| FINAL\_TABLE\_2  /X\_Reduced\_Simple | 3 | * All areas * Correlated features * No Venues * No Migrants   Features   * 'Job\_Density' * 'Median\_Income\_H' * 'Anxiety' * 'Bars#' * 'Coffee#' * 'Extra#’ * 'Suvirval\_Rate\_2Years’ * 'Population' | Maybe |  |
| FINAL\_TABLE\_2  /X\_Reduced | 4 | * All areas * Restaurants as dummies * Correlated features * No Migrants   Features   * 'Job\_Density' * 'Median\_Income\_H' * 'Anxiety' * 'Bars#','Coffee#' * 'Extra#' * 'Suvirval\_Rate\_2Years' * 'Population' * Restaurants\_Dummies | Yes |  |
| FINAL\_TABLE2\_  TO\_KMEAN | 5 | * All restaurants * All areas * Migrants * All features (except duplicated)   Features:   * 'Population' * 'Population\_Density' * 'Working\_Population' * 'Youth\_Population', * 'Elderly\_Population', * 'Born\_Abroad', * 'Employment', * 'Median\_Income\_H', * 'Job\_Density', * 'Active\_Business' * 'Suvirval\_Rate\_2Years' * 'Life\_Satisfcation' * 'Anxiety' * 'Restaurants# * 'Bars#' * 'Coffee#' * 'Extra#, * Restaurants Dummies * Migrants Dummies | NO | A close up of a map  Description automatically generated |
| FINAL\_TABLE2\_TO\_  KMEAN2 | 6 | * All areas * No Restaurants * Migrants * All features (except duplicated)   Features   * 'Population' * 'Population\_Density' * 'Working\_Population' * 'Youth\_Population' * 'Elderly\_Population' * 'Born\_Abroad', 'Employment' * 'Median\_Income\_H' * 'Suvirval\_Rate\_2Years' * 'Life\_Satisfcation', * Migrants Dummies | Maybe | A picture containing text, map  Description automatically generated |

­­Table 1: Clustering results

## Intermediary Conclusions

* Restaurants categories or restaurants will play a much smaller role than expected but it’s better than just using all venues categories
* In all simulations City of London is London is a separate cluster on its own
* The selected data set contains:
  + Correlated features to restaurants (Significant correlation factor when compared with restaurants numbers in that area)
  + Restaurants types: Italian restaurant, Burger Join, etc..

## Cluster analysis

Cluster 0

* Areas covered: Haringey
* Properties:
  + Low on restaurants
  + Relative high population density but low job density
  + Lower household income
* Conclusion
  + Not recemented to invest in restaurants in this area, seems more like an industrial area or a quiet residential area
  + Does not seem the most attractive area for restaurant’s landscape

A screenshot of a cell phone

Description automatically generated

Table 2: Restaurants distribution Cluster 0

Cluster1

* Areas covered:

'Bromley', 'Camden', 'Croydon', 'Ealing', 'Hackney', 'Havering',

'Kingston upon Thames', 'Lewisham', 'Merton', 'Wandsworth'

* Properties:
  + Relative low density
  + Relative higher elderly population
  + High employment rate
  + Better than cluster 1 median earnings
  + Highest business success rate
  + Decent number of restaurants
* Conclusion
  + Seems to be a stable attractive area
  + But does not seem to be an attractive/successful place for English restaurants

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| # | Rest\_Tpye | Count | # | Rest\_Tpye | Count |
| 1 | Burger Joint | 15 | 16 | French Restaurant | 4 |
| 2 | Italian Restaurant | 13 | 17 | Fish & Chips Shop | 3 |
| 3 | Restaurant | 8 | 18 | Ramen Restaurant | 2 |
| 4 | Asian Restaurant | 8 | 19 | American Restaurant | 2 |
| 5 | Vegetarian / Vegan Restaurant | 7 | 20 | Mexican Restaurant | 2 |
| 6 | Pizza Place | 7 | 21 | African Restaurant | 2 |
| 7 | Portuguese Restaurant | 7 | 22 | Malay Restaurant | 2 |
| 8 | Sandwich Place | 7 | 23 | Spanish Restaurant | 2 |
| 9 | Fast Food Restaurant | 7 | 24 | Middle Eastern Restaurant | 2 |
| 10 | Thai Restaurant | 6 | 25 | Falafel Restaurant | 2 |
| 11 | Vietnamese Restaurant | 6 | 26 | English Restaurant | 2 |
| 12 | Sushi Restaurant | 6 | 27 | New American Restaurant | 1 |
| 13 | Caribbean Restaurant | 5 | 28 | Japanese Restaurant | 1 |
| 14 | Turkish Restaurant | 4 | 29 | Latin American Restaurant | 1 |
| 15 | Indian Restaurant | 4 | 30 | Cajun / Creole Restaurant | 1 |
|  |  |  |  |  |  |

Table 3: Restaurants distribution Cluster 1

Cluster2

* Areas covered: 'Greenwich', 'Islington', 'Kensington and Chelsea', 'Tower Hamlets'
* Properties:
  + Highest density populated of all districts
  + Highest diversity
  + Highest job density
  + Lower elderly population suggesting a more dynamic and crowded area
  + Lower survival rate suggesting a competitive landscape
  + Considerable number of restaurants
  + More interest shown for English restaurants
  + Second highest income area
* Conclusion
  + Seems to be an attractive area with good dynamics
  + Seems to be the best place for a English restaurant

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| # | Rest\_Tpye | Count | # | Rest\_Tpye | Count |
| 1 | Burger Joint | 11 | 16 | Portuguese Restaurant | 2 |
| 2 | Sushi Restaurant | 6 | 17 | Sandwich Place | 2 |
| 3 | Pizza Place | 5 | 18 | Dumpling Restaurant | 1 |
| 4 | Japanese Restaurant | 5 | 19 | Ramen Restaurant | 1 |
| 5 | Italian Restaurant | 5 | 20 | Filipino Restaurant | 1 |
| 6 | English Restaurant | 4 | 21 | Persian Restaurant | 1 |
| 7 | Mediterranean Restaurant | 4 | 22 | Afghan Restaurant | 1 |
| 8 | French Restaurant | 3 | 23 | Vietnamese Restaurant | 1 |
| 9 | Restaurant | 3 | 24 | Kebab Restaurant | 1 |
| 10 | Mexican Restaurant | 3 | 25 | Austrian Restaurant | 1 |
| 11 | Steakhouse | 3 | 26 | Latin American Restaurant | 1 |
| 12 | Middle Eastern Restaurant | 2 | 27 | Thai Restaurant | 1 |
| 13 | Modern European Restaurant | 2 | 28 | Chinese Restaurant | 1 |
| 14 | Spanish Restaurant | 2 |  |  |  |
| 15 | Indian Restaurant | 2 |  |  |  |

Table 4: Restaurants distribution Cluster 2

Cluster3

* Areas covered: City of London
* Properties:
  + Low density suggesting a relative expensive area with limited housing
  + Covers only the city center of London and seems to be the most peculiar cluster
  + Low employment rate
  + Lower elderly population suggesting a more dynamic and crowded area
  + Very high job density
  + Competitive business landscape with the lowest success rate of business
  + High bars and restaurants density suggesting a very promising area
  + High “Extra” attraction suggest a very tourist area
  + By far the highest median income area
* Conclusion
  + Probably the most attractive area to invest with high potential but a risky one non the less.
  + Expected to have high competition, cost of operation and initial cost
  + English restaurants seems not be very prominent, but nevertheless a good restaurant in this area can be very successful

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| # | Rest\_Tpye | Count | # | Rest\_Tpye | Count |
| 1 | Steakhouse | 3 | 10 | Asian Restaurant | 2 |
| 2 | Restaurant | 3 | 11 | English Restaurant | 1 |
| 3 | French Restaurant | 3 | 12 | Mexican Restaurant | 1 |
| 4 | Italian Restaurant | 3 | 13 | Falafel Restaurant | 1 |
| 5 | Seafood Restaurant | 3 | 14 | Scandinavian Restaurant | 1 |
| 6 | Modern European Restaurant | 2 | 15 | Udon Restaurant | 1 |
| 7 | Sushi Restaurant | 2 | 16 | Latin American Restaurant | 1 |
| 8 | Indian Restaurant | 2 | 17 | New American Restaurant | 1 |
| 9 | Vietnamese Restaurant | 2 | 18 | Pizza Place | 1 |

Table 5: Restaurants distribution Cluster 3

Cluster4

* Areas covered Barnet, Bexley, Brent, Enfield, Hammersmith and Fulham,Harrow, Hillingdon, Hounslow, Lambeth, Newham,Redbridge, Southwark, Sutton
* Properties
  + Median population density with a relative lower number of available restaurants
  + Relative low median income and lower job density than the rest of clusters
  + Fairly stable business continuity
* Conclusion
  + Can be an interesting opportunity for investment for a small restaurant with no high income expectations
  + Overall not the most attractive cluster to invest

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| # | Rest\_Tpye | Count | # | Rest\_Tpye | Count |
| 1 | Sandwich Place | 16 | 18 | Tapas Restaurant | 2 |
| 2 | Fast Food Restaurant | 15 | 19 | Sushi Restaurant | 2 |
| 3 | Italian Restaurant | 11 | 20 | Vegetarian / Vegan Restaurant | 2 |
| 4 | Indian Restaurant | 11 | 21 | Portuguese Restaurant | 2 |
| 5 | Chinese Restaurant | 8 | 22 | Japanese Restaurant | 1 |
| 6 | Restaurant | 8 | 23 | Romanian Restaurant | 1 |
| 7 | Pizza Place | 7 | 24 | Dim Sum Restaurant | 1 |
| 8 | Burger Joint | 7 | 25 | French Restaurant | 1 |
| 9 | Turkish Restaurant | 6 | 26 | Latin American Restaurant | 1 |
| 10 | English Restaurant | 3 | 27 | Afghan Restaurant | 1 |
| 11 | Fish & Chips Shop | 3 | 28 | Israeli Restaurant | 1 |
| 12 | Steakhouse | 3 | 29 | Greek Restaurant | 1 |
| 13 | Asian Restaurant | 3 | 30 | Vietnamese Restaurant | 1 |
| 14 | Korean Restaurant | 3 | 31 | Argentinian Restaurant | 1 |
| 15 | Ramen Restaurant | 2 | 32 | Modern European Restaurant | 1 |
| 16 | Eastern European Restaurant | 2 | 33 | Mexican Restaurant | 1 |
| 17 | Thai Restaurant | 2 | 34 | Spanish Restaurant | 1 |

Table 6: Restaurants distribution Cluster 4

A screenshot of a cell phone

Description automatically generated

Table 7: Overall mean values for all five clusters

# Final conclusions

* There seems to be a lack of English restaurants resulting from the data collected using Foursquare, placing the English restaurants category outside of top ten categories after Italian, Burgers joins, Asian restaurants, Indian Restaurants and others
* Surprisingly the number of French, Portuguese or Vietnamese restaurants exceeds the English ones (this might be due to the limited data collection), nevertheless the 1400+ venues collected should be a significant sampler for making the above assumptions
* Multiple analysis scenarios were performed where different features were selected, and the results are presented in the table xx above. Out of these scenarios one that seems to have the better cluster distribution was selected. Scenario 4 was selected
* The analysis concluded above allow us to segregate the London restaurants landscape into 5 clusters based on restaurants types and numbers and the data was completed with other significant information like: population density, job density, household income, working population, business survival rate, life satisfaction, anxiety and coffee shops and bars numbers
* Thought the analysis the City of London itself was always placed in a separate individual cluster making us realize the uniquness of this area
* However the most promising scenario are Clusters 2 ('Greenwich', 'Islington', 'Kensington and Chelsea', 'Tower Hamlets') and Cluster 3 (City of London) showing high potential for new restaurants. Out of these two scenario the safest bet seems to be on Cluster 2 because it seems less riskier and challenging than Cluster 3. See detailed analysis and conclusins on the above chaptes

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| # | Rest\_Tpye | Count | # | Rest\_Tpye | Count |
| 1 | Italian Restaurant | 33 | 29 | Seafood Restaurant | 4 |
| 2 | Burger Joint | 33 | 30 | Korean Restaurant | 4 |
| 3 | Sandwich Place | 25 | 31 | Latin American Restaurant | 4 |
| 4 | Fast Food Restaurant | 23 | 32 | African Restaurant | 3 |
| 5 | Restaurant | 22 | 33 | Greek Restaurant | 2 |
| 6 | Pizza Place | 20 | 34 | Polish Restaurant | 2 |
| 7 | Indian Restaurant | 20 | 35 | Eastern European Restaurant | 2 |
| 8 | Sushi Restaurant | 16 | 36 | Afghan Restaurant | 2 |
| 9 | Asian Restaurant | 13 | 37 | Tapas Restaurant | 2 |
| 10 | Turkish Restaurant | 11 | 38 | Malay Restaurant | 2 |
| 11 | Portuguese Restaurant | 11 | 39 | American Restaurant | 2 |
| 12 | French Restaurant | 11 | 40 | Kebab Restaurant | 2 |
| 13 | Vietnamese Restaurant | 10 | 41 | New American Restaurant | 2 |
| 14 | English Restaurant | 10 | 42 | Southern / Soul Food Restaurant | 1 |
| 15 | Vegetarian / Vegan Restaurant | 9 | 43 | Cajun / Creole Restaurant | 1 |
| 16 | Chinese Restaurant | 9 | 44 | Argentinian Restaurant | 1 |
| 17 | Thai Restaurant | 9 | 45 | Israeli Restaurant | 1 |
| 18 | Steakhouse | 9 | 46 | Udon Restaurant | 1 |
| 19 | Mexican Restaurant | 7 | 47 | Brazilian Restaurant | 1 |
| 20 | Japanese Restaurant | 7 | 48 | Dim Sum Restaurant | 1 |
| 21 | Mediterranean Restaurant | 6 | 49 | Persian Restaurant | 1 |
| 22 | Fish & Chips Shop | 6 | 50 | German Restaurant | 1 |
| 23 | Caribbean Restaurant | 5 | 51 | Filipino Restaurant | 1 |
| 24 | Ramen Restaurant | 5 | 52 | Austrian Restaurant | 1 |
| 25 | Spanish Restaurant | 5 | 53 | Romanian Restaurant | 1 |
| 26 | Middle Eastern Restaurant | 5 | 54 | Bulgarian Restaurant | 1 |
| 27 | Modern European Restaurant | 5 | 55 | Scandinavian Restaurant | 1 |
| 28 | Falafel Restaurant | 4 | 56 | Dumpling Restaurant |  |

Table 7: Restaurants distribution for all areas

# Future investigations

* In addition to the current analysis I see the need of:
  + Extending the data set that will require another level of account in Foursquare. The ability to add more data should results in better results
  + Add ranking for restaurants. The current analysis was performed without ranking due to limitations:
    - imposed by the Foresquare account (providing a limited number of rankings)
    - Inability to use web scraping techniques due to protection from Foresquare
  + Changing the account type should allow the collection of rating, but it comes with additional costs
  + The current analysis concluded that Cluster 2 is the best candidate, however the investigation can continue inside Cluster 2. The same type of analysis can be performed in order to detect which particular section of this Cluster is most suited for opening a restaurant. Adding cost of rent in the equation might also increase the output of the analysis