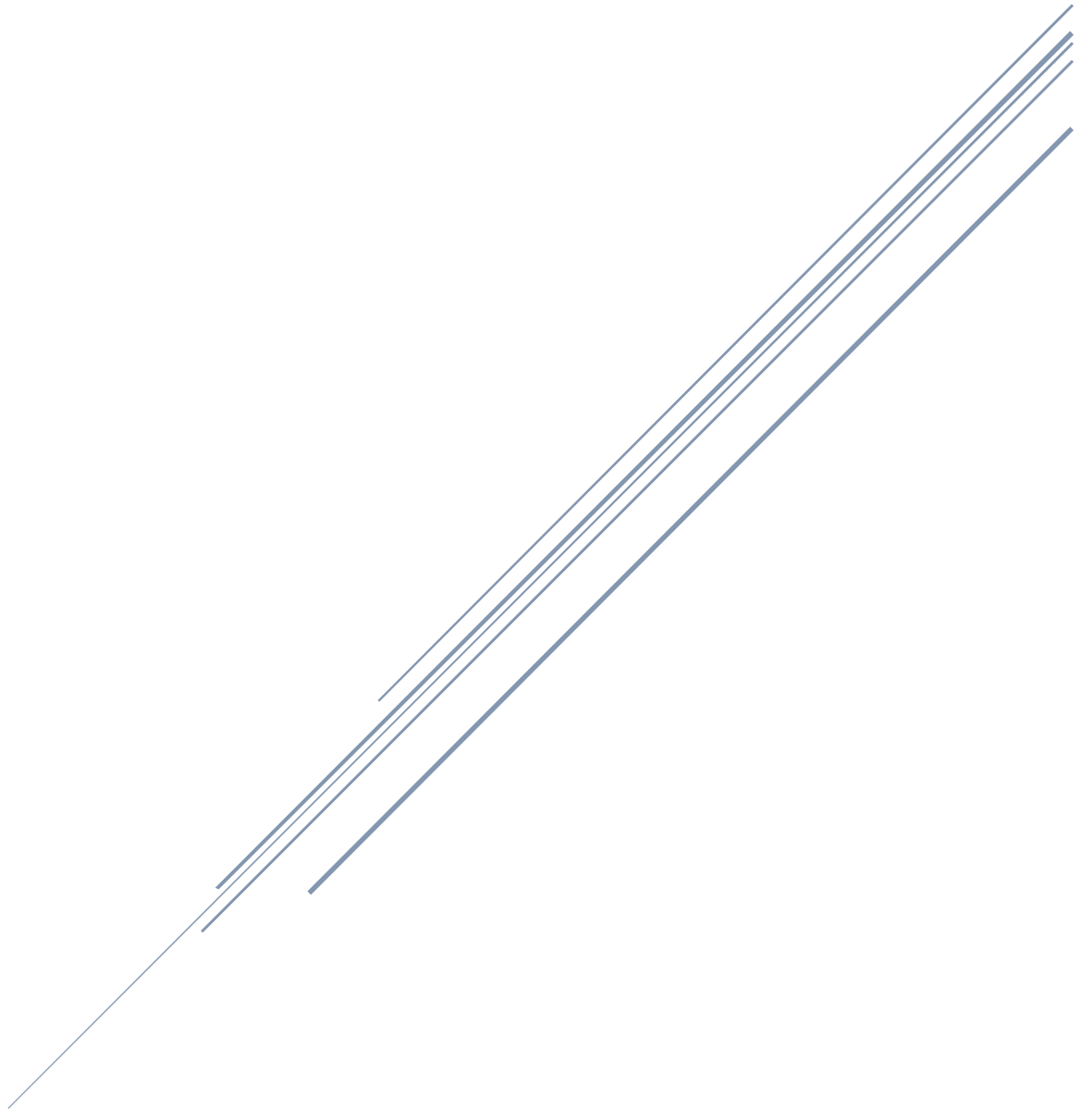


ANALYSING THE MOST SUITABLE LONDON AREA FOR OPENING AN ENGLISH RESTAURANT

Capstone Project - ML



Sebastian Moldovan

Table of Contents

EXECUTIVE SUMMARY.....	2
BACKGROUND AND PROBLEM.....	2
INTEREST.....	3
DATA ACQUISITION AND DATA CLEANUP.....	3
DATA CLEANING	3
CLASSIFICATION	5
INTERMEDIARY CONCLUSIONS.....	10
CLUSTER ANALYSIS	10
FINAL CONCLUSIONS	16
FUTURE INVESTIGATIONS	18

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 places 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
 - o 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

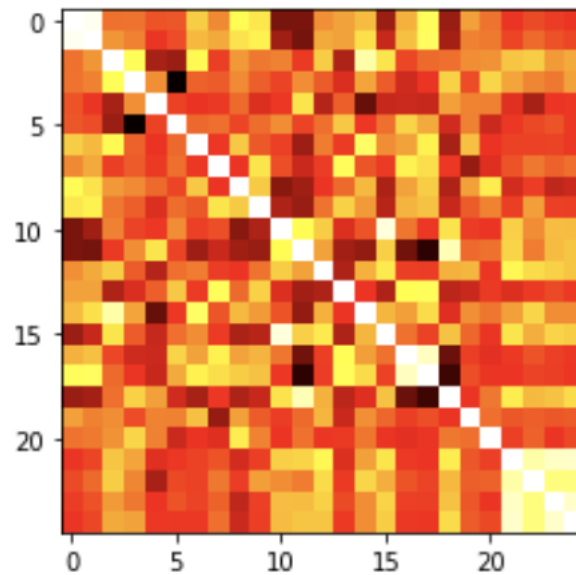


Figure 1: Correlation of the features

- 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#', 'Survival_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)

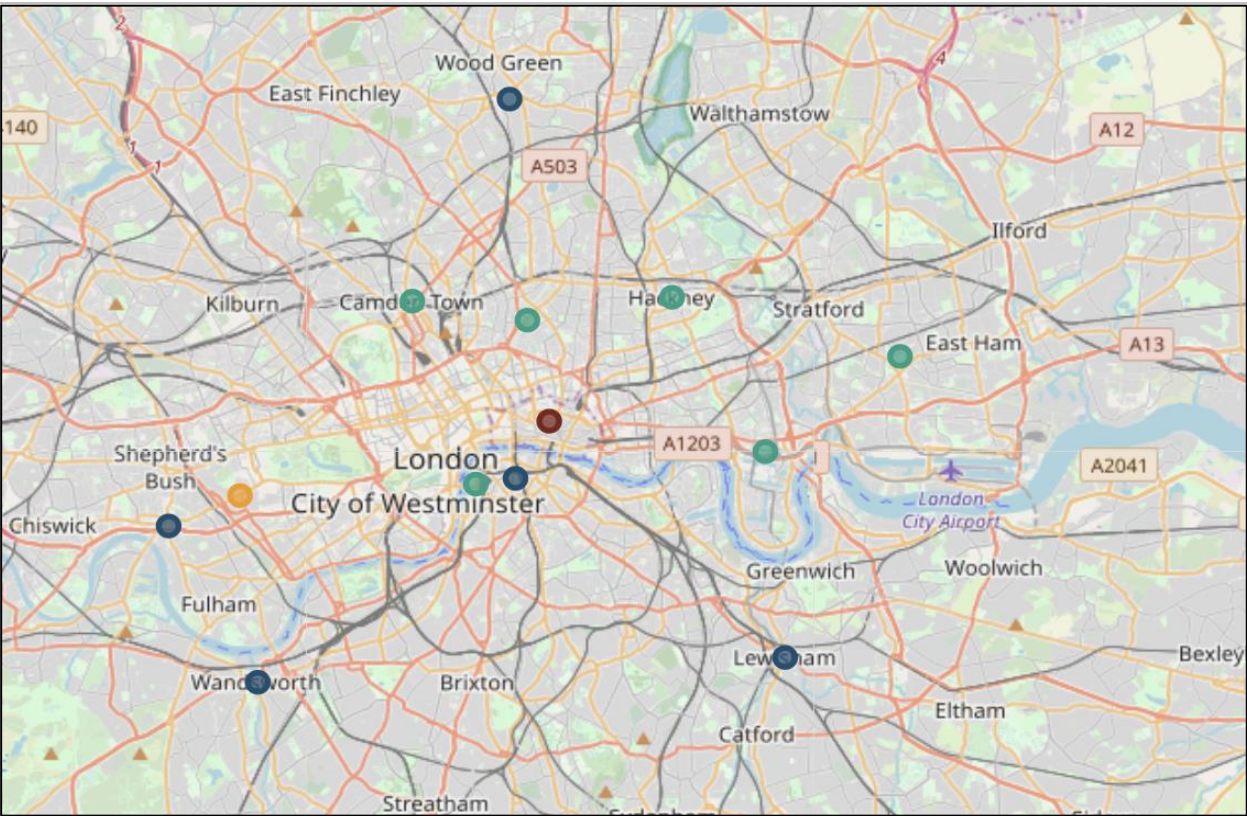
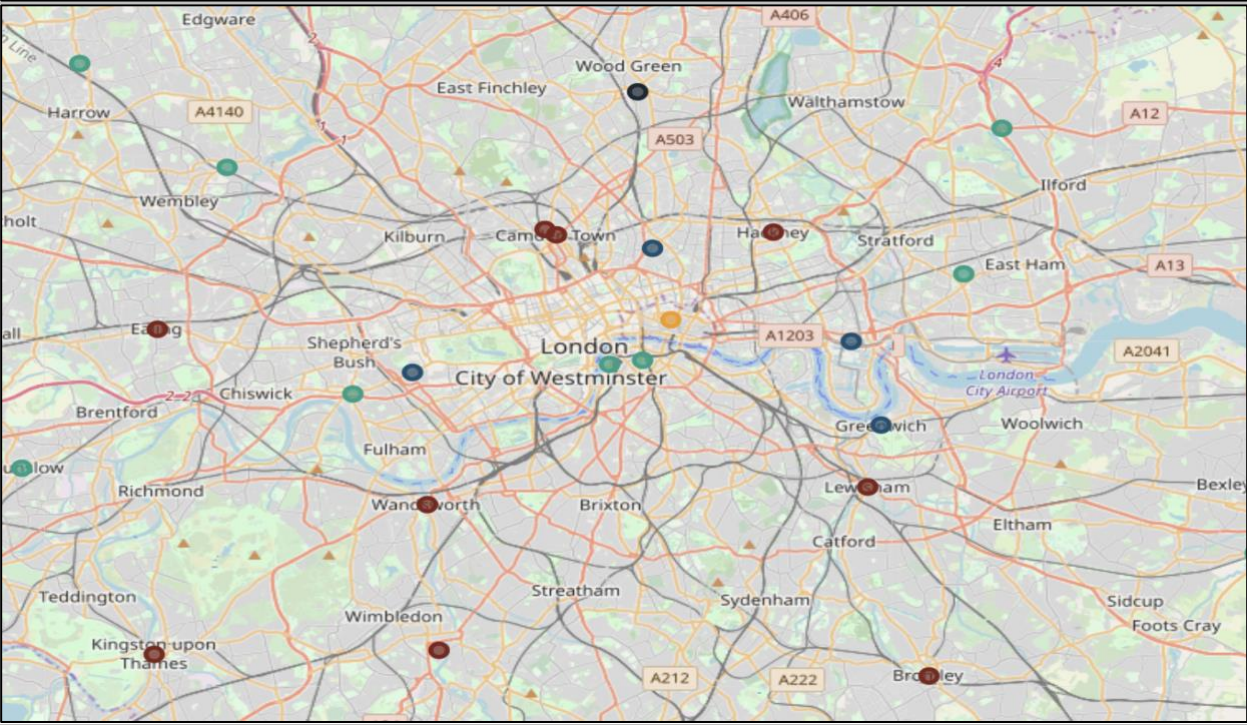
Classification

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 $\text{SRQT}(N/2)$ 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	<div><div><div>- Inner London</div><div>- All Venues</div><div>- No Migrants</div></div><div>Features</div><div><div>- 'Population'</div><div>- 'Households',</div><div>- 'Population_Density'</div><div>- 'Working_Population'</div><div>- 'Youth_Population',</div><div>- 'Elderly_Population'</div><div>- 'Born_Abroad'</div><div>- 'Employment'</div><div>- 'Unemployment'</div><div>- 'Annual_Pay'</div><div>- 'Median_Income_H'</div><div>- 'Job_Density', 'Active_Business'</div><div>- 'Survival_Rate_2Years'</div><div>- 'Crime'</div><div>- 'House_Price'</div><div>- 'Life_Satisfcation'</div><div>- 'Happiness'</div><div>- 'Anxiety'</div><div>- Top10 Venues</div></div></div>	Maybe	

Maybe

A / X_A	2	<ul style="list-style-type: none">- Inner London- No Venues- No Migrants <p>Features</p> <ul style="list-style-type: none">- 'Population'- 'Population_Density',- 'Working_Population',- Youth_Population'- 'Elderly_Population',- Born_Abroad',- 'Employment',- Median_Income_H',- Job_Density'- 'Active_Business',- 'Survival_Rate_2Years'- 'Life_Satisfaction'- 'Anxiety'	NO	
FINAL_TABLE_2 /X_Reduced_Simple	3	<ul style="list-style-type: none">- All areas- Correlated features- No Venues- No Migrants <p>Features</p> <ul style="list-style-type: none">- 'Job_Density'- 'Median_Income_H'- 'Anxiety'- 'Bars#'- 'Coffee#'- 'Extra#'- 'Survival_Rate_2Years'- 'Population'	Maybe	

FINAL_TABLE_2 /X_Reduced	4	<ul style="list-style-type: none">- All areas- Restaurants as dummies- Correlated features- No Migrants <p>Features</p> <ul style="list-style-type: none">- 'Job_Density'- 'Median_Income_H'- 'Anxiety'- 'Bars#','Coffee#'- 'Extra#'- 'Survival_Rate_2Years'- 'Population'- Restaurants_Dummies	Yes	
---------------------------------	---	---	-----	--

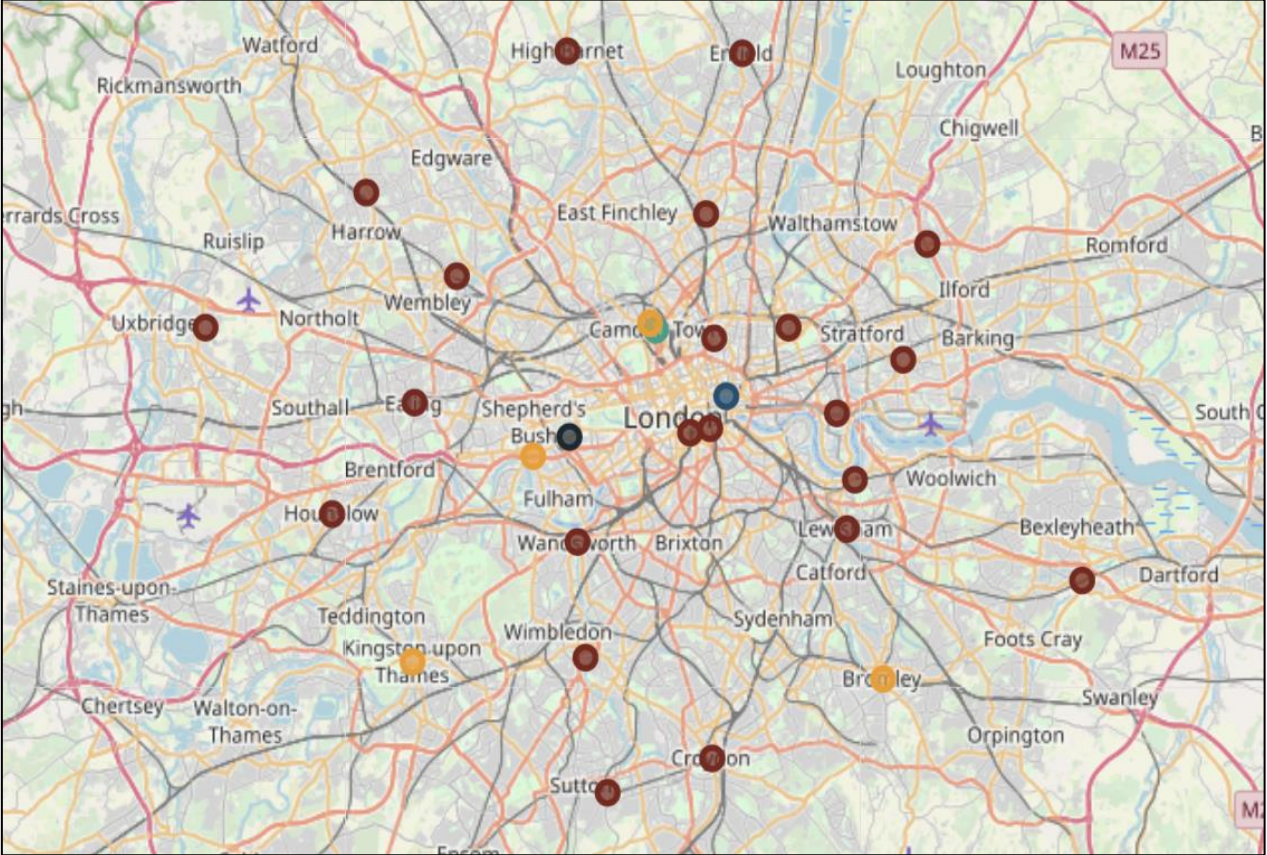
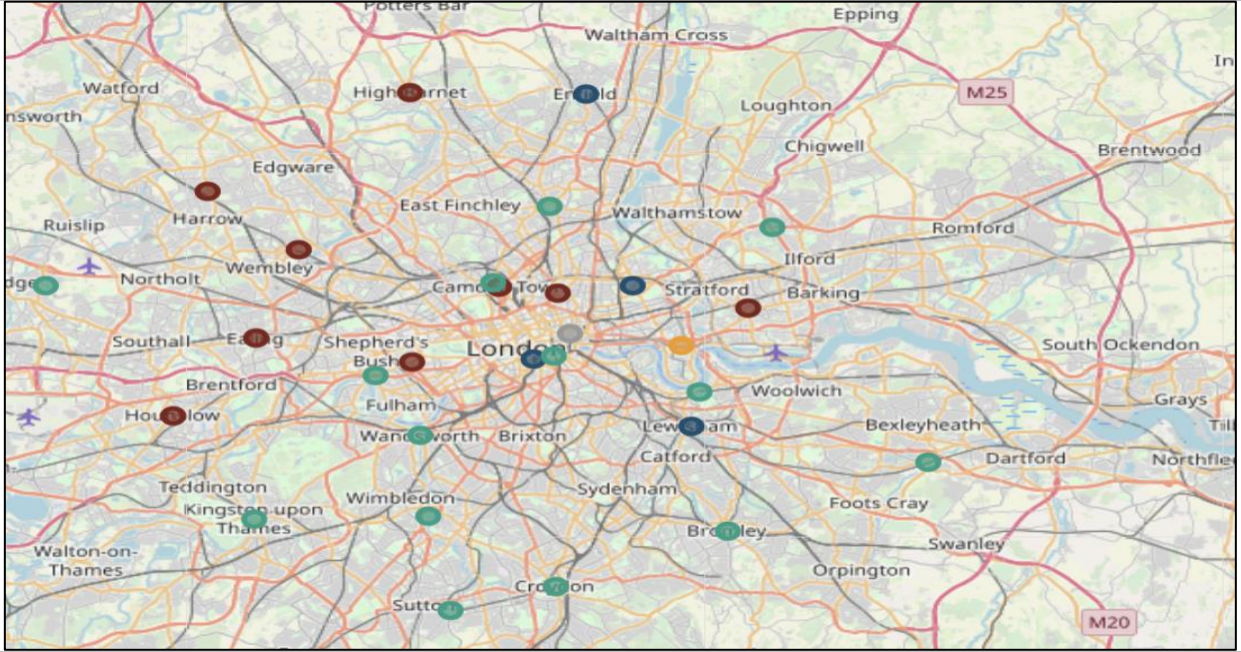
FINAL_TABLE2_ TO_KMEAN	5	<ul style="list-style-type: none"> - All restaurants - All areas - Migrants - All features (except duplicated) <p>Features:</p> <ul style="list-style-type: none"> - 'Population' - 'Population_Density' - 'Working_Population' - 'Youth_Population', - 'Elderly_Population', - 'Born_Abroad', - 'Employment', - 'Median_Income_H', - 'Job_Density', - 'Active_Business' - 'Suvirval_Rate_2Years' - 'Life_Satisfcation' - 'Anxiety' - 'Restaurants#' - 'Bars#' - 'Coffee#' - 'Extra#', - Restaurants Dummies - Migrants Dummies 	NO	
FINAL_TABLE2_TO_ KMEAN2	6	<ul style="list-style-type: none"> - All areas - No Restaurants - Migrants - All features (except duplicated) <p>Features</p> <ul style="list-style-type: none"> - 'Population' - 'Population_Density' - 'Working_Population' - 'Youth_Population' - 'Elderly_Population' - 'Born_Abroad', 'Employment' - 'Median_Income_H' - 'Suvirval_Rate_2Years' - 'Life_Satisfcation', - Migrants Dummies 	Maybe	

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:
 - o Correlated features to restaurants (Significant correlation factor when compared with restaurants numbers in that area)
 - o Restaurants types: Italian restaurant, Burger Join, etc.

Cluster analysis

Cluster 0

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

	Rest_Tpye	Count
0	Indian Restaurant	1
1	Mediterranean Restaurant	1
2	Italian Restaurant	1
3	Polish Restaurant	1
4	African Restaurant	1
5	Middle Eastern Restaurant	1
6	Fast Food Restaurant	1
7	Bulgarian Restaurant	1
8	Falafel Restaurant	1
9	Turkish Restaurant	1

Table 2: Restaurants distribution Cluster 0

Cluster1

- Areas covered:
 - 'Bromley', 'Camden', 'Croydon', 'Ealing', 'Hackney', 'Havering',
 - 'Kingston upon Thames', 'Lewisham', 'Merton', 'Wandsworth'
- Properties:
 - o Relative low density
 - o Relative higher elderly population
 - o High employment rate
 - o Better than cluster 1 median earnings
 - o Highest business success rate
 - o Decent number of restaurants
- Conclusion
 - o Seems to be a stable attractive area
 - o 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:
 - o Highest density populated of all districts
 - o Highest diversity
 - o Highest job density
 - o Lower elderly population suggesting a more dynamic and crowded area
 - o Lower survival rate suggesting a competitive landscape
 - o Considerable number of restaurants
 - o 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 an 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
 - o Median population density with a relative lower number of available restaurants
 - o Relatively low median income and lower job density than the rest of clusters
 - o Fairly stable business continuity
- Conclusion
 - o Can be an interesting opportunity for investment for a small restaurant with no high income expectations
 - o 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

	0	1	2	3	4
Population	274803.00	282068.300	240129.2500	8548.00	289986.769231
Population_Density	92.70	68.330	122.7000	28.90	66.176923
Working_Population	90.70	87.690	90.1000	90.60	87.900000
Youth_Population	19.50	18.540	18.3250	27.20	20.446154
Elderly_Population	9.30	12.310	9.9000	9.40	12.100000
Born_Abroad	39.60	31.960	40.5500	0.00	38.607692
Employment	71.30	74.600	70.8250	64.60	73.100000
Median_Income_H	45860.00	52472.000	65347.5000	99390.00	46666.923077
Job_Density	0.48	0.793	1.1425	84.60	0.760769
Active_Business	11875.00	14807.500	14466.2500	19250.00	13952.692308
Survival_Rate_2Years	71.00	74.600	69.5000	63.00	72.000000
Life_Satisfaction	7.24	7.266	7.3000	6.59	7.330769
Anxiety	3.18	3.465	3.3650	5.57	3.150769
Restaurants#	10.00	14.700	18.7500	33.00	10.153846
Bars#	2.00	6.800	6.7500	7.00	4.307692
Coffee#	5.00	7.700	5.2500	12.00	4.230769
Extra#	15.00	26.300	31.7500	48.00	16.000000

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 joints, 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 uniqueness 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 conclusions on the above chapters

#	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	1

Table 7: Restaurants distribution for all areas

Future investigations

- In addition to the current analysis I see the need of:
 - o Extending the data set that will require another level of account in Foursquare. The ability to add more data should results in better results
 - o Add ranking for restaurants. The current analysis was performed without ranking due to limitations:
 - imposed by the Foursquare account (providing a limited number of rankings)
 - Inability to use web scraping techniques due to protection from Foursquare
 - o 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