SHORTLISTING OPTIMUM LOCATIONS FOR AN ITALIAN RESTAURANT IN HAMBURG

Faizan Ahmed

11.08.2020

Background/ Business Problem

A person wants to open an Italian restaurant in Hamburg. Finding a suitable spot for Italian restaurant is particularly challenging because there are many Italian restaurants in Hamburg. Therefore, he wants to use data science and machine learning techniques to find a suitable location for his restaurant. Restaurant should ideally be located in part of Hamburg where the owner could get a significant customer base. He is targeting the mid to upper segment of the market so this we will be looking for areas with high buying power and less competition.

There are many neighborhoods in Hamburg some more affluent then the others and the population density varies widely between the different parts of the city. The aim of this project is to shortlist suitable locations considering these factors.

Description of Data

The Data sources used for this project are as follows:

1. A list of all the neighborhoods in Hamburg including their population density is taken from the Wikipedia page: https://de.wikipedia.org/wiki/Liste_der_Bezirke_und_Stadtteile_Hamburgs
Web scraping is used to read the relevant data on this webpage.

2. The data on the average income in each neighborhood is taken from the report this website:
https://www.statistik-nord.de/fileadmin/Dokumente/Statistik_informiert_SPEZIAL/SI_SPEZIAL_VIII_2017.pdf
A csv file is generated using the data in this pdf and read into pandas dataframe.

3. The data on venues in each neighborhood serving food is obtained using the Foursquare places API.

From these sources, data on the population density, average income and the number and type of food venues will be used to shortlist potential locations for an Italian restaurant.

Methodology section

This section will be devoted to execution of all the steps required in this project. This section will consist of:

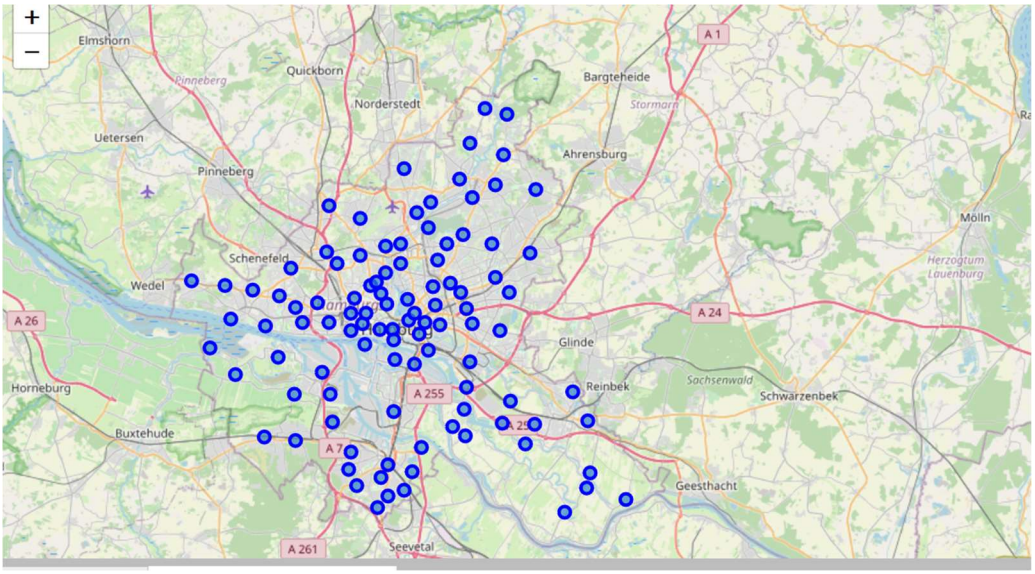
1. Data Collection
2. Exploring Data
3. Preprocessing data for analysis
4. Clustering data using K-Means Clustering

1. Data Collection

Data on the neighborhoods is collected using web scraping from Wikipedia. Beautiful soup library from python is used for web scraping. The coordinates for each location are obtained using Nominatim.

	Neighborhood	Borough	Population	Population_per_km2	latitude	longitude
0	Hamburg-Altstadt	Hamburg-Mitte	2350	979	53.550468	9.994640
1	HafenCity	Hamburg-Mitte	4925	2239	53.542913	9.995835
2	Neustadt	Hamburg-Mitte	12762	5549	53.549881	9.979048
3	St. Pauli	Hamburg-Mitte	22097	8839	53.553935	9.959432
4	St. Georg	Hamburg-Mitte	11358	4733	53.556993	10.014162

The map of Hamburg is shown below with all the neighborhoods marked as blue dots. Python library folium is used to visualize the map.



The data on average income for each neighborhood is obtained from source [2]. This data is then added to the dataframe of neighborhood. Any missing income data is filled with the average of the column.

	Neighborhood	Borough	Population	Population_per_km2	latitude	longitude	Average income
0	Hamburg-Altstadt	Hamburg-Mitte	2350	979	53.550468	9.994640	31336.0
1	HafenCity	Hamburg-Mitte	4925	2239	53.542913	9.995835	93206.0
2	Neustadt	Hamburg-Mitte	12762	5549	53.549881	9.979048	34521.0
3	St. Pauli	Hamburg-Mitte	22097	8839	53.553935	9.959432	27977.0
4	St. Georg	Hamburg-Mitte	11358	4733	53.556993	10.014162	44121.0

The Data on food venues is obtained using the Foursquare API.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Hamburg-Altstadt	53.550468	9.99464	Café Paris	53.550106	9.994227	Café
1	Hamburg-Altstadt	53.550468	9.99464	frittenwerk	53.551136	9.994500	Fast Food Restaurant
2	Hamburg-Altstadt	53.550468	9.99464	Pastaria Da Franco	53.551148	9.994428	Italian Restaurant
3	Hamburg-Altstadt	53.550468	9.99464	Salam-City	53.549939	9.995543	Falafel Restaurant
4	Hamburg-Altstadt	53.550468	9.99464	Picasso	53.549934	9.995627	Spanish Restaurant

2. Exploring Data

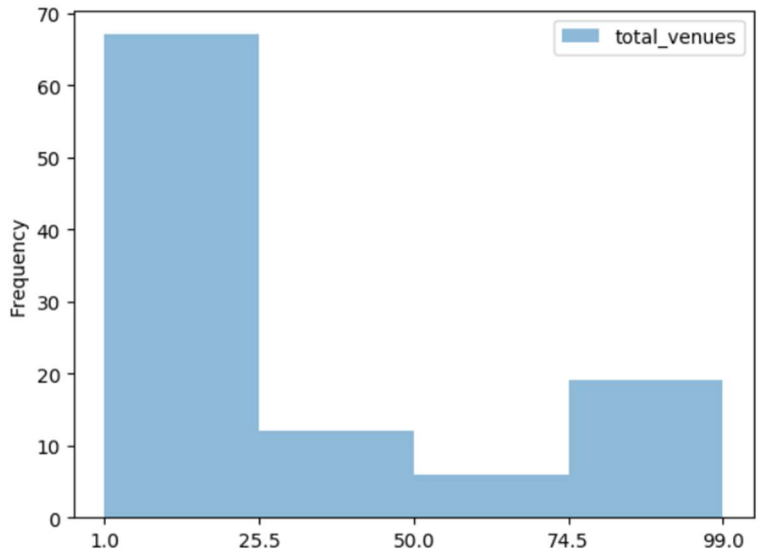
In this section, we will take a first look at the data collected. This is important to get some insight into data or possibly delete or add more data for our modelling requirements.

The data collected from Foursquare is grouped according to venue category.

Venue Category	
Bakery	402
Café	399
Italian Restaurant	242
German Restaurant	226
Restaurant	186

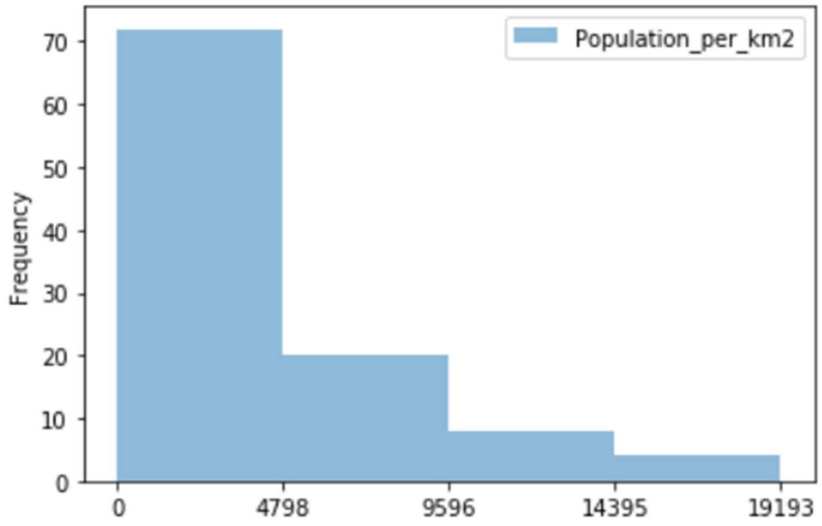
We see here that bakeries are the most common type of venue. Since bakeries are not in direct competitors of a restaurant this category is deleted from the dataframe for later analysis. Another information we see here is that Italians restaurants are even more common than German restaurants so finding the optimum location is even more critical for the success of Italian restaurant.

Using the data on food venues and grouping it based on neighborhoods we can get the total number of venues in each neighborhood. This data can be shown in the form of a histogram below.



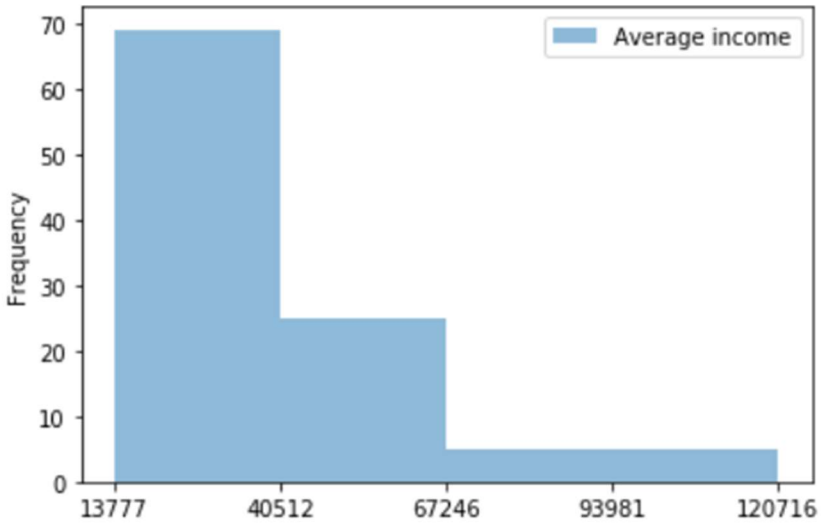
We see that most of the neighborhoods have under 25 food venues. These bins can be classified as **low, medium, high and very high competition.**

To get insight into population density a histogram is plotted as shown below.



The data for this histogram is divided into four bins. These bins can be classified as **low/average, medium, high and very high population density**. The plot shows that most of the neighborhoods have a density of under 5000 residents per square km.

Similarly, a histogram for average income is shown below.



Here the bins can also be classified as **low, medium, high and very high income.**

3. Preprocessing data for analysis

A view of dataframe containing the neighborhoods and the numerical features to be used for KMeans Clustering is shown below

	Neighborhood	Population_per_km2	Average income	total_venues
0	Hamburg-Altstadt	979	31336.0	97
1	HafenCity	2239	93206.0	93
2	Neustadt	5549	34521.0	99
3	St. Pauli	8839	27977.0	98
4	St. Georg	4733	44121.0	94

The features vary greatly in range and must be standardized for the clustering algorithm to give meaningful results.

StandardScaler() function is used from the python scikit-learn library for this purpose. The standardized features are as follows:

	Neighborhood	Population_per_km2	Average income	total_venues
0	Hamburg-Altstadt	-0.726399	-0.480244	2.042690
1	HafenCity	-0.434408	2.418739	1.922328
2	Neustadt	0.332648	-0.331008	2.102871
3	St. Pauli	1.095069	-0.637634	2.072781
4	St. Georg	0.143549	0.118810	1.952418

To help in the decision making process after clustering the top 10 most common venues for each neighborhood are found from the data from Foursquare API

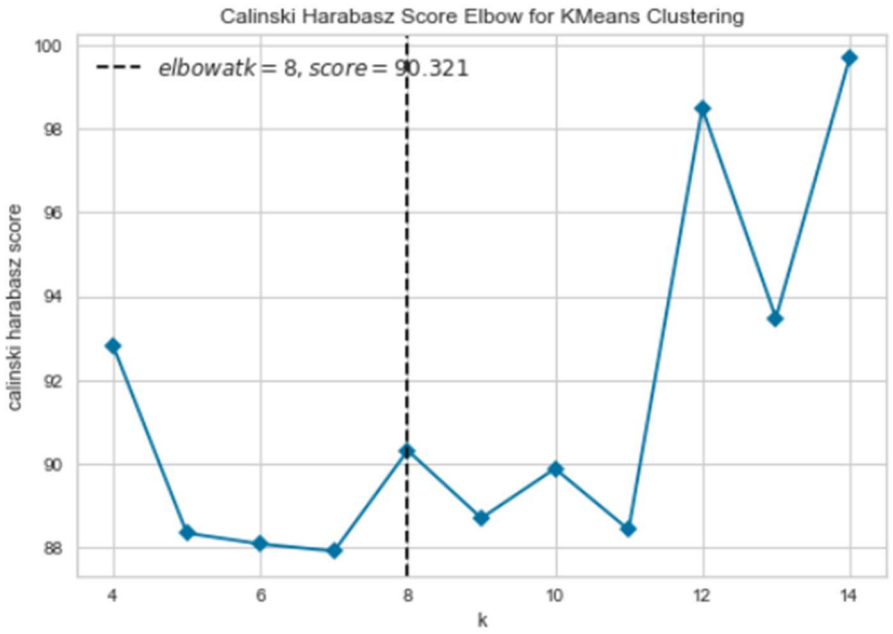
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Allermöhe	German Restaurant	Wings Joint	Food Truck	Dumpling Restaurant	Eastern European Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant	Fish & Chips Shop
1	Alsterdorf	Restaurant	Café	Vietnamese Restaurant	Asian Restaurant	Greek Restaurant	Doner Restaurant	Burger Joint	German Restaurant	Lebanese Restaurant	Fast Food Restaurant
2	Altengamme	Diner	Café	Food Truck	Eastern European Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant	Fish & Chips Shop	Food
3	Altenwerder	Diner	Gastropub	Food Truck	Dumpling Restaurant	Eastern European Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant	Fish & Chips Shop
4	Altona-Altstadt	Café	Seafood Restaurant	Italian Restaurant	Restaurant	Pizza Place	German Restaurant	Burger Joint	French Restaurant	Asian Restaurant	Thai Restaurant
5	Altona-Nord	Café	Pizza Place	Burger Joint	German Restaurant	Asian Restaurant	Indian Restaurant	Vietnamese Restaurant	Restaurant	Italian Restaurant	French Restaurant
6	Bahrenfeld	Restaurant	Italian Restaurant	German Restaurant	Café	Snack Place	Indian Restaurant	BBQ Joint	Asian Restaurant	Diner	Fast Food Restaurant
7	Barmbek-Nord	Café	Italian Restaurant	Taverna	German Restaurant	Fast Food Restaurant	Falafel Restaurant	Greek Restaurant	Asian Restaurant	Thai Restaurant	Indian Restaurant
8	Barmbek-Süd	Café	German Restaurant	Doner Restaurant	Italian Restaurant	Indian Restaurant	Fast Food Restaurant	Pizza Place	Greek Restaurant	Taverna	Vietnamese Restaurant
9	Bergedorf	Restaurant	German Restaurant	Falafel Restaurant	Fast Food Restaurant	Café	Burger Joint	Asian Restaurant	Trattoria/Osteria	Sushi Restaurant	Steakhouse

4. Clustering data using K-Means clustering

To find the shortlist optimum locations for the restaurant. The machine-learning algorithm K-Means clustering is used to cluster the neighborhoods in Hamburg according to the three selected characteristics i.e population density, average income and the total number of venues. K-Means clustering is a type of unsupervised learning.

This algorithm is particularly suitable in this scenario because will make the decision making process easier by clustering neighborhoods with similar characteristics.

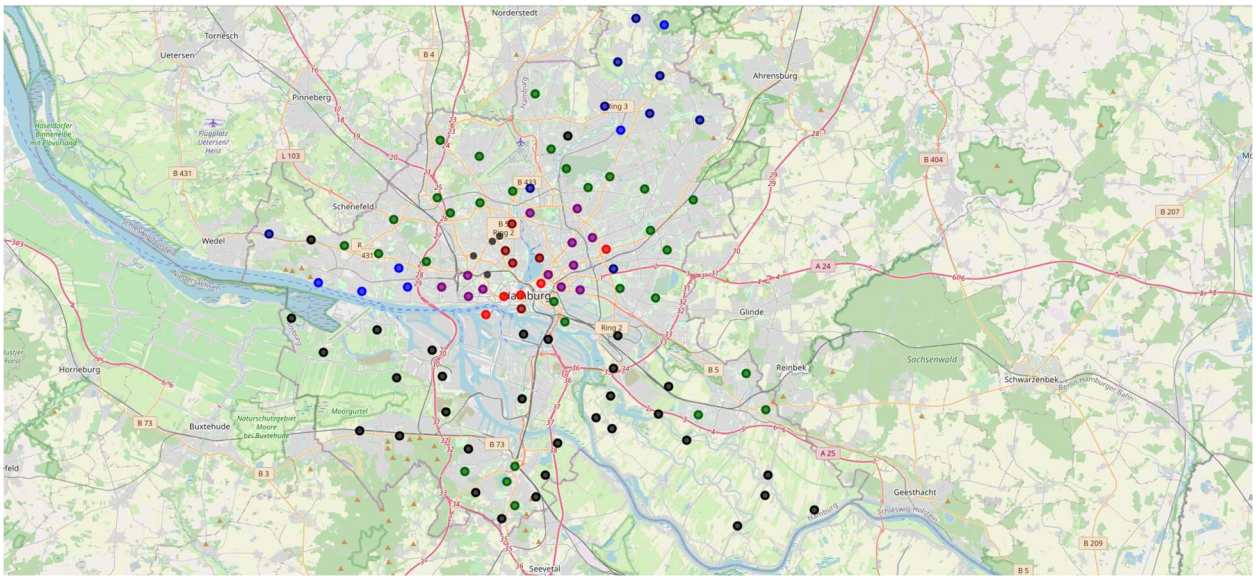
The data for clustering is already scaled using the StandardScaler() function in the last section. The next question is what value of the number of clusters (K) is optimum for out data. Elbow method will be used to find the optimum value of K. The KElbowVisualizer() function in Yellowbrick library is used to implement this method. This graph below shows the 'calinski harabasz' score which is calculated for each K and the optimum value of K found by the function is also shown as **K=8**.



Next K-Means clustering is performed and the cluster labels are added to the neighborhoods along with the features used for clustering and extra data that can help us make better decision. A slice of the dataframe is shown below.

	Neighborhood	Population	Population_per_km2	Average income	total_venues	Cluster Labels	latitude	longitude	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	Hamburg-Altstadt	2350	979	31336.0	97	3	53.550468	9.994640	German Restaurant	Italian Restaurant	Asian Restaurant	Café	Steakhouse	Vietnamese Restaurant	Burger Joint	Restaurant	Seafood Restaurant	French Restaurant
1	HafenCity	4925	2239	93206.0	93	7	53.542913	9.995835	German Restaurant	Italian Restaurant	Café	Restaurant	Asian Restaurant	Bistro	Seafood Restaurant	Burger Joint	Vietnamese Restaurant	French Restaurant
2	Neustadt	12762	5549	34521.0	99	3	53.549881	9.979048	Café	German Restaurant	Italian Restaurant	Seafood Restaurant	Steakhouse	French Restaurant	Trattoria/Osteria	Restaurant	Austrian Restaurant	Burger Joint
3	St. Pauli	22097	8839	27977.0	98	0	53.553935	9.959432	Café	Seafood Restaurant	Restaurant	German Restaurant	Pizza Place	Burger Joint	Asian Restaurant	Gastropub	Austrian Restaurant	Italian Restaurant
4	St. Georg	11358	4733	44121.0	94	3	53.556993	10.014162	Italian Restaurant	Café	Restaurant	German Restaurant	Vietnamese Restaurant	Burger Joint	Asian Restaurant	French Restaurant	Mediterranean Restaurant	Thai Restaurant

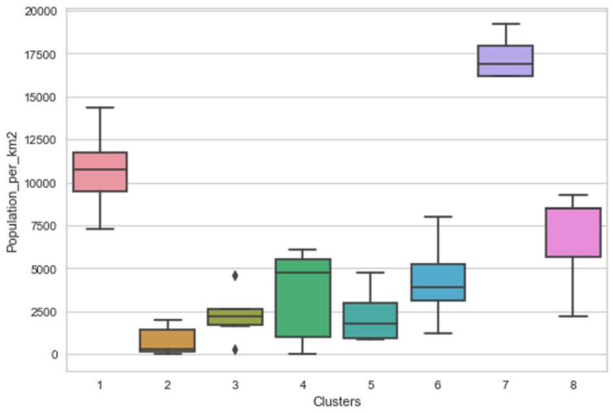
A map of Hamburg with neighborhoods represented using different colored markers is plotted to get a visual representation of the clusters

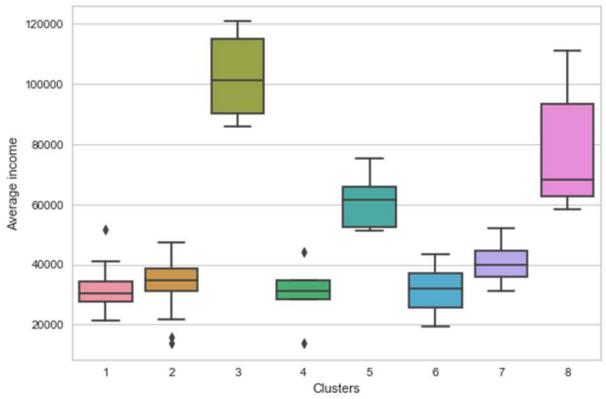


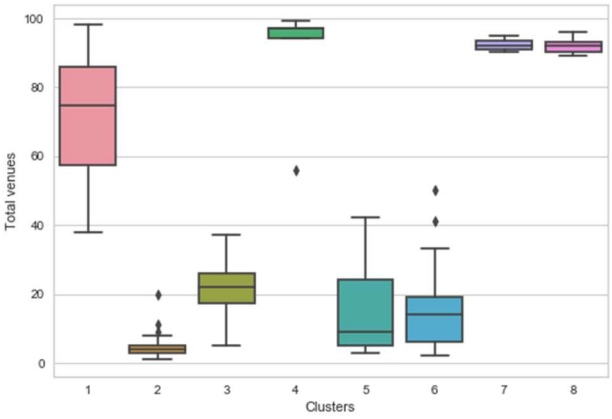
Results and Discussion

In this section, we will look at the results we have obtained from clustering and try to get an insight into those results.

Let us have a look at the characteristics of each cluster







Looking at the above comparison one can see that that clusters 1, 2, 4, 6 and 7 are towards the lower end of the income spectrum therefore these areas would not be very suitable for opening a restaurant targeting the mid to upper level segment of the market.

Cluster 8 includes areas with high average income but looking at the number of venues there, this area has very high competition. The last two remaining clusters 3 and 5 seem promising.

Cluster 3

	Neighborhood	Population	Population_per_km2	Average income	total_venues	Cluster Labels	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
24	Groß Flottbek	11111	4630	85952.0	37	2	Café	Italian Restaurant	Restaurant	Indian Restaurant	Asian Restaurant	German Restaurant	Sushi Restaurant	Snack Place	Doner Restaurant	Portuguese Restaurant
25	Othmarschen	15737	2623	108258.0	18	2	Café	German Restaurant	Restaurant	Burger Joint	Falafel Restaurant	Deli / Bodega	Steakhouse	Sushi Restaurant	Italian Restaurant	Brazilian Restaurant
28	Nienstedten	7181	1670	120716.0	17	2	Café	Seafood Restaurant	Snack Place	Restaurant	German Restaurant	French Restaurant	Portuguese Restaurant	Diner	Fried Chicken Joint	Ethiopian Restaurant
29	Blankenese	13730	1783	117139.0	26	2	Café	Seafood Restaurant	Restaurant	Snack Place	Italian Restaurant	French Restaurant	Fast Food Restaurant	Steakhouse	Sushi Restaurant	Burger Joint
63	Wellingsbüttel	10848	2646	88606.0	26	2	Italian Restaurant	Café	Sushi Restaurant	Trattoria/Osteria	Seafood Restaurant	Pizza Place	Eastern European Restaurant	Restaurant	Sandwich Place	Burger Joint
69	Wohldorf-Ohlstedt	4650	269	94234.0	5	2	German Restaurant	Wings Joint	Food Truck	Dumpling Restaurant	Eastern European Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant	Fish & Chips Shop

Cluster 5

	Neighborhood	Population	Population_per_km2	Average income	total_venues	Cluster Labels	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
32	Rissen	15886	951	65855.0	3	4	Indian Restaurant	German Restaurant	Café	Wings Joint	Food Court	Eastern European Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant
45	Alsterdorf	15204	4751	52426.0	27	4	Restaurant	Café	Vietnamese Restaurant	Asian Restaurant	Greek Restaurant	Doner Restaurant	Burger Joint	German Restaurant	Lebanese Restaurant	Fast Food Restaurant
57	Marienthal	13521	4225	59131.0	42	4	Chinese Restaurant	Asian Restaurant	Café	Turkish Restaurant	Italian Restaurant	Doner Restaurant	Burger Joint	Seafood Restaurant	Vietnamese Restaurant	Dumpling Restaurant
64	Sasel	23750	2827	61360.0	9	4	Asian Restaurant	Indian Restaurant	Trattoria/Osteria	Taverna	Chinese Restaurant	Gastropub	Sushi Restaurant	Food	Food Court	Eastern European Restaurant
65	Poppenbüttel	23901	2951	52157.0	24	4	Italian Restaurant	Restaurant	Fast Food Restaurant	Seafood Restaurant	Café	Sushi Restaurant	Deli / Bodega	Sandwich Place	Middle Eastern Restaurant	Steakhouse
67	Lemsahl-Mellingstedt	6852	857	75191.0	5	4	German Restaurant	Restaurant	Pizza Place	Seafood Restaurant	Food	Dumpling Restaurant	Eastern European Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant
68	Duvenstedt	6160	906	65694.0	5	4	Café	Pizza Place	German Restaurant	Modern European Restaurant	Burger Joint	Wings Joint	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant
70	Bergstedt	10736	1534	51374.0	6	4	German Restaurant	Italian Restaurant	Snack Place	Pizza Place	Restaurant	BBQ Joint	Food Truck	Eastern European Restaurant	English Restaurant	Ethiopian Restaurant
71	Volksdorf	20978	1808	63763.0	17	4	Café	Greek Restaurant	Italian Restaurant	Tapas Restaurant	Fast Food Restaurant	Seafood Restaurant	Doner Restaurant	Brazilian Restaurant	Steakhouse	German Restaurant

The neighborhoods included in these clusters have a good income and the competition is not that high. The population density is on the lower side but with sufficient buying power, the restaurant can still do good business with regular customers and high profit margins.

Out of the clusters 3 and 5 the five neighborhoods that seem the promising based on the gathered data are shortlisted are shown below.

	Neighborhood	Population	Population_per_km2	Average income	total_venues	Cluster Labels	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
28	Nienstedten	7181	1670	120716.0	17	2	Café	Seafood Restaurant	Snack Place	Restaurant	German Restaurant	French Restaurant	Portuguese Restaurant	Diner	Fried Chicken Joint	Ethiopian Restaurant
69	Wohldorf-Ohlstedt	4650	269	94234.0	5	2	German Restaurant	Wings Joint	Food Truck	Dumpling Restaurant	Eastern European Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant	Fish & Chips Shop
32	Rissen	15886	951	65855.0	3	4	Indian Restaurant	German Restaurant	Café	Wings Joint	Food Court	Eastern European Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant
67	Lemsahl-Mellingstedt	6852	857	75191.0	5	4	German Restaurant	Restaurant	Pizza Place	Seafood Restaurant	Food	Dumpling Restaurant	Eastern European Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant
68	Duvenstedt	6160	906	65694.0	5	4	Café	Pizza Place	German Restaurant	Modern European Restaurant	Burger Joint	Wings Joint	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant

It can be seen that although population density is not high in these areas but people living in these areas have a high buying power. Moreover, the competition is low and only those neighborhoods are selected where there are no Italian restaurants in the top 10 food venues.

Conclusion

It is shown through this project that data science and machine learning can be used to make effective business decisions. Machine learning models like K-Means clustering can be used to get insights into data by using a large amount of data, which would otherwise not be possible just by looking at it.

In this project neighborhoods of Hamburg are analyzed to shortlist optimum locations for opening an Italian restaurant catering to the mid to upper end segment of the market. Taking the targeted customers into consideration locations are shortlisted using data on food venues from foursquare API, and information on average income and population density available on the internet.

Five neighborhoods are shortlisted where the population has high buying power and the competition is low. The stakeholder may consider other factors like real estate availability, running cost, availability of workforce etc. before making a final decision.