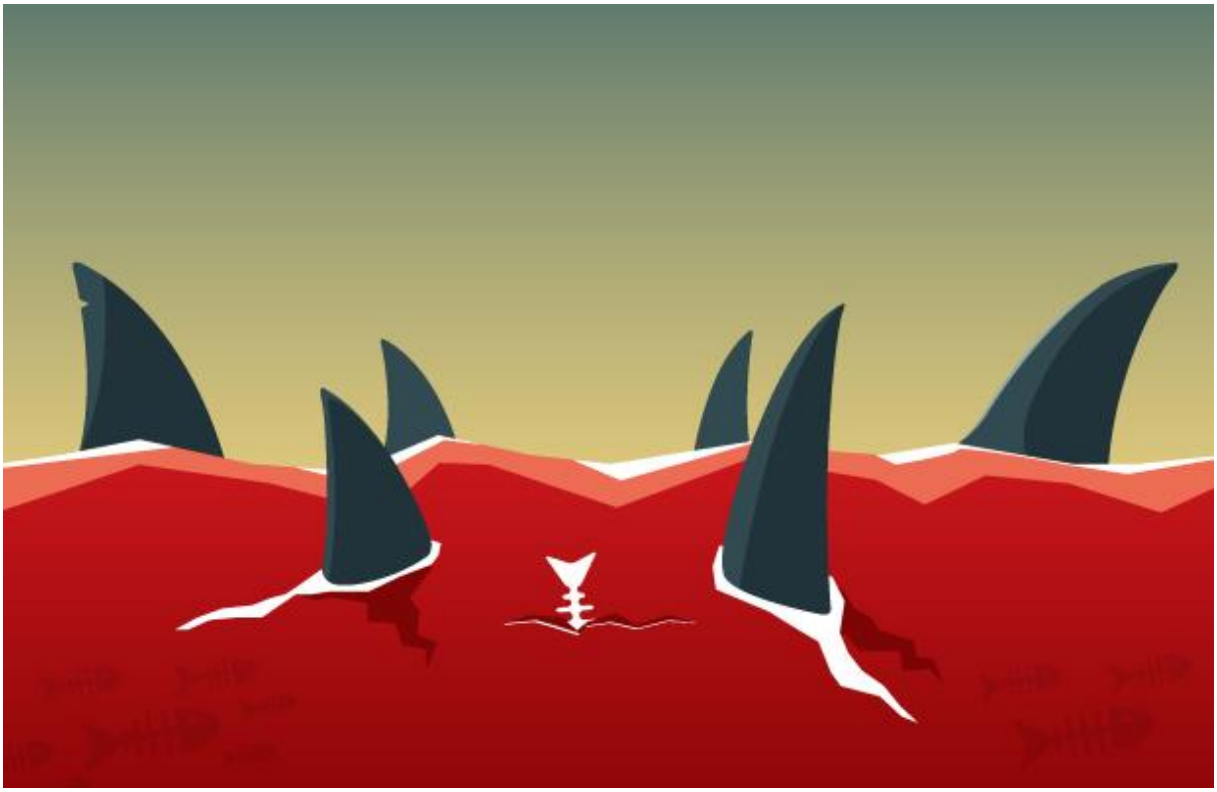
Red Ocean Locator



Business Problem

1. Description of the project

For a group of entrepreneurs who would like to start their own restaurant, the location is critical. From our analysis, the perfect location would be a place where there is medium to high density of restaurant. But instead of competing head to head with same restaurant type, the strategy is to make the competition irrelevant simply by opening up a service which does not exist yet. This approach is directly derived from the blue ocean marketing strategy.

The concept is quite simple to understand. The Red Ocean is where every restaurant is today. There is a defined market, defined competitors and a typical way to run restaurant. The Red Ocean is analogous to a shark infested ocean where the sharks are fighting each other for the same prey. The Blue Ocean, on the other hand, is calm, smooth and little or no competition. The goal is to capture some of the existing customers but to attract new ones. And the service offered (beyond the food) is something none of the restaurant are currently offering. Simply because they do not know about it or they do not know how to do it. Of course, they will try to copy but it will take some time. In this capstone case, we will not cover what kind of services these new types of restaurant will offer. However, since they want to be located to in high density restaurant neighbourhood, our red ocean locator will provide entrepreneurs a list of neighbourhoods forming the red ocean.

2. Data

Geo-locational information about all the city neighbourhood will be required. We specifically and technically mean the latitude and longitude numbers. The city selected for this project will be Amsterdam as the entrepreneurs are dutch and wishes to identify a few places to start up new restaurants. This data will be provided by ourselves since entrepreneurs do not have this information. Neighborhoods in Amsterdam will be identitifed by their corresponding Postal Codes.

Data about different venues including restaurants in different neighborhoods will also be required. In order to get this dataset, "Foursquare" will be queried. The output will return basic and advanced information about each venue. For example, venue data will contain its precise latitude, longitude distance to the neighbourhood center but also advanced information such as the category of that venue. A typical request from Foursquare will provide us with the following information:

[Postal Code] [Venue Latitude] [Venue Longitude] [Venue name] [Venue Category] [Distance (meter)]

3. Background information about Foursquare :

Foursquare is a local search-and-discovery service mobile app which provides search results for its users. The app provides personalized recommendations of places to go to near a user's current location based on users' "previous browsing history, purchases, or check-in history"

Community: More than 50 million people use Foursquare City Guide and Foursquare Swarm each month, across desktop, mobile web, and mobile apps. We recently surpassed more than 12 billion check-ins, and our record high is over 9 million check-ins in a single day on Swarm.
Platform: More than 105 million venues mapped around the world.
Employees: Nearly 250 people at our New York headquarters or based in San Francisco, Chicago, Los Angeles, Atlanta, Detroit, London, Singapore and more.

Data Section

1. Identifying Amsterdam Neighborhoods

We will use Postal Codes of different neighbourhoods in Amsterdam to find the list of neighborhoods. We will essentially obtain our information from https://maps.amsterdam.nl/?LANG=en and then process the tables inside this site. Collected data from this site shows postcode, neighbourhood, surface m2, polygons coordinates, longitude and latitude. The figure (Fig.1) hereunder shows partially all the data collected.

	Postal Code	Neighbourhood	Longitude	Latitude
0	1011	Amsterdam-Noord	4.866168	52.422284
1	1012	Petroleum	4.843031	52.410937
2	1013	De Pijp	4.890788	52.354114
3	1014	Zuidelijke	4.891925	52.363208
4	1015	oostelijk Havengebied	4.935600	52.374593
5	1016	Burgwallen	4.897117	52.374927
6	1017	Westelijke Grachtengordel	4.881871	52.370431
7	1018	Haarlemmerbuurt	4.874058	52.396327
8	1019	Alfa Driehoek	4.862575	52.393100
9	1021	Staatsliedenbuurt	4.870957	52.379134
10	1022	Westerlijke Markanaal	4.854686	52.373668
11	1023	Oud Oost	4.929871	52.360946
12	1024	Indisch Buurt (westelijk deel/0	4.936080	52.362187
13	1025	Watergraafsmeer	4.948973	52.349637

Fig.1 – Postal Code and Neghbourhood

After data selection, a map of Amsterdam can be built using Folium library to show all neighbourhoof of Amsterdam. We end up with 54 neighbourhoods in Amsterdam. The figure (Fig.2) hereunder shows all the postal code with corresponding neighbourhoods.

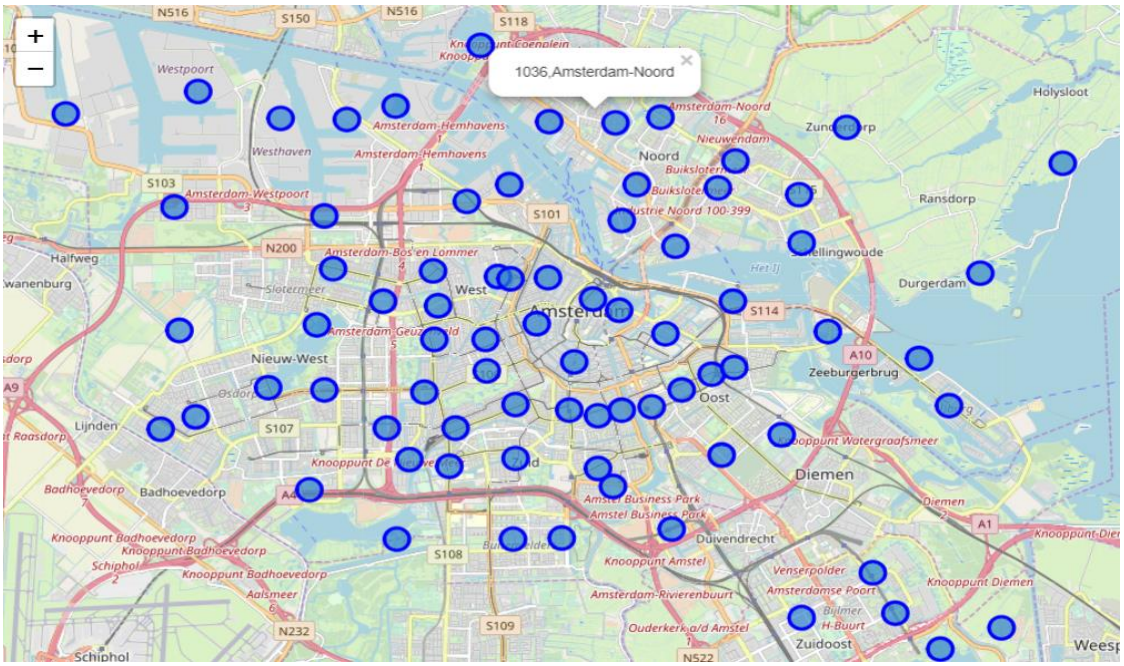


Fig.2 – Amsterdam Postal Code and Neighbourhoods

2. Connecting to Foursquare and Retrieving Venue Data for Each Neighborhood

Once the list of neighborhoods is completed, a request to the Foursquare API is made to gather information about venues inside each and every neighborhood. The default radius is the neighbourhood boundary. It means that Foursquare returns to all venues located in the neighborhood..) We end up with a total of 2096 venues and 265 unique categories. The figure (Fig.3) hereunder shows partially all the data collected.

Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Postal Code	Venue	Venue Latitude	Venue Longitude	Venue Category
Afrikahaven	3	3	3	3	3	3	3
Alfa Driehoek	9	9	9	9	9	9	9
Amsterdam-Noord	9	9	9	9	9	9	9
Apollobuurt	24	24	24	24	24	24	24
Australienhaven	5	5	5	5	5	5	5
Bijlmer Centrum	42	42	42	42	42	42	42
Bijlmer oost	16	16	16	16	16	16	16
Bloemenbuurt	21	21	21	21	21	21	21
Bos en Lommer	45	45	45	45	45	45	45
Buiksloot	13	13	13	13	13	13	13
Buitenveldert	54	54	54	54	54	54	54
Bullewijk	15	15	15	15	15	15	15
Burgwallen	100	100	100	100	100	100	100
De Pijp	100	100	100	100	100	100	100
Driemond	2	2	2	2	2	2	2

Fig.3 – Venues count per neighbourhood

3. Preparing Data and Creating a Dataframe for the all the restaurant Venues in Amsterdam

Since we want to focus only on restaurant to identify red ocean, we will filter out all the categories which do not point of refer to restaurant. This brings down the number of venues identified as restaurant to 499 in Amsterdam distributed into 52 restaurant categories. At this stage, every restaurant is identified by its name, category, longitude, latitude as well as postal code, neighbourhood, neighbourhood latitude and longitude (8 column).

After this stage, the venue category (52 restaurant categories) gets One-hot encoded After On-hot encoding, a new dataframe is created to group all the restaurant per neighbourhood with an extra column displaying the total restaurant per neighbourhood.

Now, the dataset is fully ready to be used for machine learning (and statistical analysis) purposes.
The figure (Fig.4) hereunder shows partially all the data collected.

	Neighbourhood	African Restaurant	American Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	Belgian Restaurant	Brazilian Restaurant	Caribbean Restaurant	Chinese Restaurant	...	Spanish Restaurant	Sushi Restaurant
0	Alfa Driehoek	1	0	0	0	0	0	0	0	0	...	0	0
1	Apollobuurt	0	0	0	0	0	0	0	0	1	...	0	0
2	Bijlmer Centrum	0	0	0	1	0	0	0	0	3	...	0	0
3	Bijlmer oost	0	0	0	0	0	0	0	0	0	...	0	0
4	Bloemenbuurt	0	0	0	0	0	0	0	0	0	...	0	0
5	Bos en Lommer	0	0	0	0	0	0	0	0	0	...	0	1
6	Buiksloot	0	0	0	0	0	0	0	0	0	...	0	0
7	Buitenvelder	0	0	0	1	0	0	0	0	0	...	0	1
8	Bullewijk	0	0	0	0	0	0	0	0	0	...	0	0
9	Burgwallen	0	1	1	0	0	0	0	0	2	...	0	1
10	De Pijp	0	0	0	0	0	0	0	0	1	...	0	1

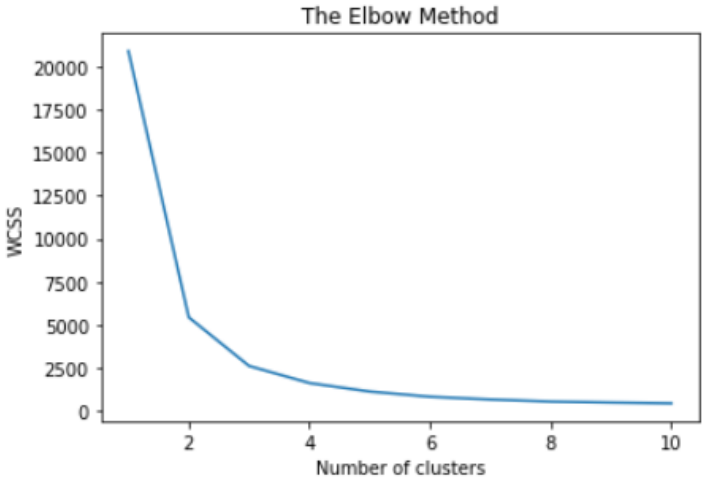
Fig.4 – one-hot encoded venues per neighbourhood

4. Applying K-Means Clustering

Before we start clustering our data, we use the silhouette score and the elbow method to determine the ideal number of clusters. Two clusters is indicated as the most efficient number but we decide to go for three for two reasons. The first one is that the Within Cluster Sum Square (WCSS) and the silhouette score of both 2 and 3 clusters is close to each other. And the second reason is that it allows us to have three levels of neighbourhood : High Red Ocean (high restaurant density), Medium Red Ocean (medium restaurant density) and Low Red Ocean (low restaurant density). The figure (Fig.5 and Fig. 6) hereunder shows the silhouette and WCSS analysis

```
For n_clusters = 2 The average silhouette_score is : 0.678711854766
For n_clusters = 3 The average silhouette_score is : 0.665225998619
For n_clusters = 4 The average silhouette_score is : 0.531350504897
For n_clusters = 5 The average silhouette_score is : 0.50745956705
For n_clusters = 6 The average silhouette_score is : 0.504539027699
For n_clusters = 7 The average silhouette_score is : 0.488517971035
For n_clusters = 8 The average silhouette_score is : 0.379760394617
For n_clusters = 9 The average silhouette_score is : 0.325113026218
```

Fig. 5 – Kmeans Silhouette score



. Fig. 6 – Kmeans WCSS score

The conclusion is that 3 clusters is enough and will cover the complexity of our problem This completes the decision on cluster number.

Before running the clustering algorithm, the only non numeric column (Neighbourhood) is removed.

After clustering we update our dataset by adding the neighbourhood column and the cluster label column representing the cluster for each neighborhood.

	Neighbourhood	Total Restaurants	Cluster Labels	Neighbourhood Latitude	Neighbourhood Longitude	Postal Code	Venue Latitude	Venue Longitude
0	Alfa Driehoek	4	0	52.393100	4.862575	1019.0	52.392864	4.868835
1	Apollobuurt	6	0	52.345157	4.875791	1079.0	52.343777	4.879163
2	Bijlmer Centrum	26	0	52.315310	4.954521	1063.0	52.314077	4.952569
3	Bijlmer oost	6	0	52.323871	4.974214	1068.0	52.322652	4.974041
4	Bloemenbuurt	4	0	52.396291	4.909152	1091.0	52.395349	4.913958

Fig.7 – Clustered Neighbourhood with Total Restaurant

5. Clustering Results and Decision making

At this stage, clusters can be visualized on a map with different colours. You consider Amsterdam to be a big circle with its downtown in the center. Cluster 1 (purple) and Cluster 2 (Blue) shows mostly restaurant in the downtown area while cluster 0 (red) shows restaurant mostly located between the outer edge of the downtown and the city limit. See Fig.8 hereunder :

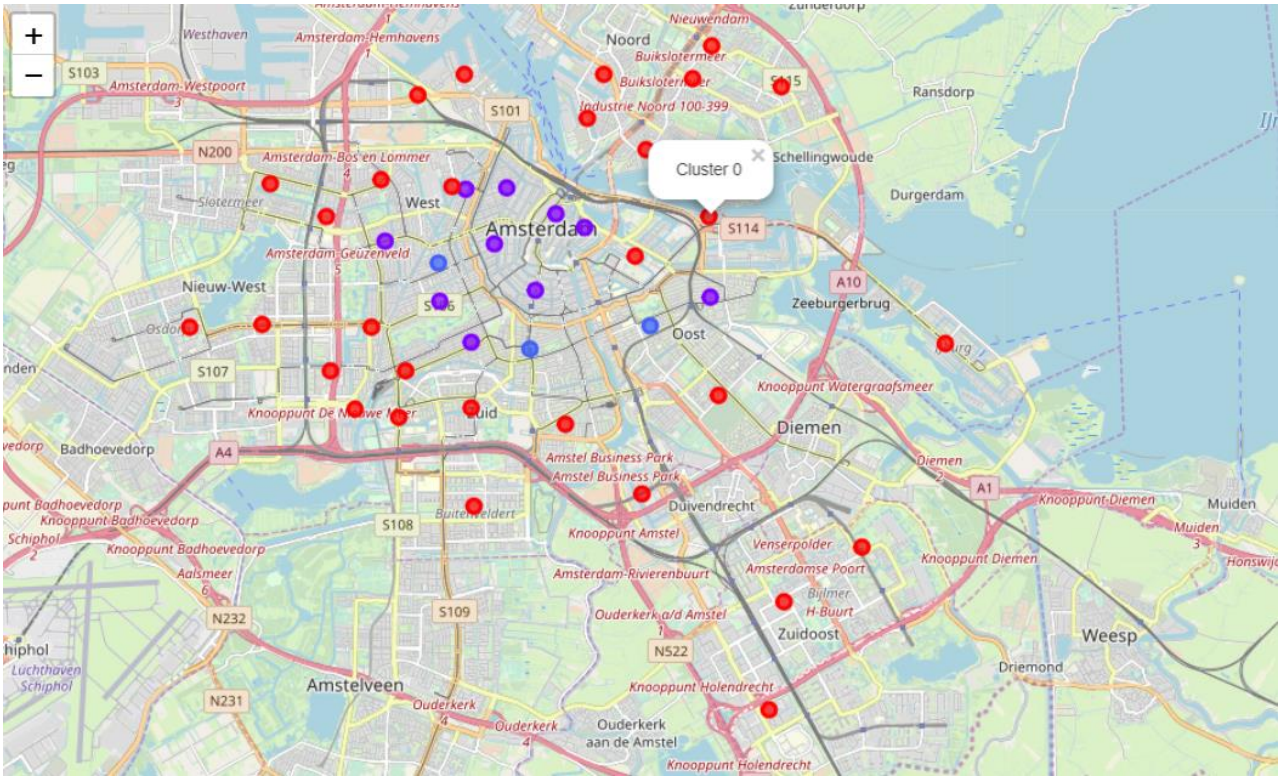


Fig.8 – Clusters of Neighbourhood grouped by restaurant density

Looking further into each cluster, we can identify each neighborhood and the total number of restaurants. For instance, in the high red ocean (Blue Cluster), three neighbourhoods shows a total comprised between 70 and 90 restaurants. See Fig.9.

Total Restaurants	Cluster Labels	Neighbourhood Latitude	Neighbourhood Longitude	Posta
74	2	52.354114	4.890788	1013.
70	2	52.367471	4.867486	1056.
90	2	52.357768	4.920982	1026.

Fig.9 – Blue Cluster - Neighbourhood grouped by restaurant density

Second best is the medium red ocean (Purple Cluster) with a total of 10 neighbourhoods showing a count of restaurant between 34 and 56 restaurants. See Fig. 10.

Total Restaurants	Cluster Labels	Neighbourhood Latitude	Neighbourhood Longitude	Posta
34	1	52.374927	4.897117	1016.(
32	1	52.378638	4.874474	1057.(
56	1	52.378982	4.884757	1055.(
40	1	52.362187	4.936080	1024.(
52	1	52.372901	4.904438	1047.(
42	1	52.355256	4.875884	1078.(
52	1	52.361583	4.868109	1033.(
46	1	52.370431	4.881871	1017.(
34	1	52.370733	4.854170	1039.
46	1	52.363208	4.891925	1014.(

Fig.10 – Purple Cluster - Neighbourhood grouped by restaurant density

Based on the machine learning output, we would advise these investors to look for a location in De Pijp, Kinkerbuurt and Oud Oost (Cluster #2) where there is the highest concentration of restaurants and then of course to choose for a different type of food service offering … to go into the blue ocean !