IBM Applied Data Science Capstone : ‘The Battle of the Neighbourhoods’

Analysis using location data via Foursquare API

Mohammad Saman

23/02/2020

Table of contents:

1. Introduction :

As the commercial capital of the Netherlands and one of the top financial centres of Europe, *Amsterdam* is known for its international atmosphere. Amsterdam is also one of the world's most multicultural cities, with at least 177 nationalities. It is home to close to half a million expatriates, with the majority of them from other parts of Europe. Due to its vibrant and diverse nightlife among other things, Amsterdam hosts close to 16.5 million visitors every year from all around the world.

Now, having stayed in Amsterdam for about a year and a half, I found quite a lot of different ethnic restaurants. Also adding to the fact that Amsterdam is a major multicultural city. As an Indian with a fondness for Italian food, I have explored the streets of Amsterdam looking for Italian restaurants and there were only a few to choose from in the city centre. The restaurants can range from being luxurious and expensive to being ordinary and affordable.

The purpose of this project is to use Foursquare location data and regional clustering of venue information to determine what might be the best neighbourhood to open a restaurant. When looking to open a business of your own, there are many variables to consider such as Location, competition, Rental costs etc. As part of this project, we are going to do an analysis of the neighbourhoods of Amsterdam for the consideration of opening an Italian restaurant. This project would be of interest to people who are looking to find suitable neighbourhoods closer to high footfall areas with reasonable rent options to open a restaurant.

2. Data Section :

For this project we need data about :

- Neighbourhood boundaries of Amsterdam
- Venues/count of venues in different neighbourhoods
- Commercial Rented space for restaurants

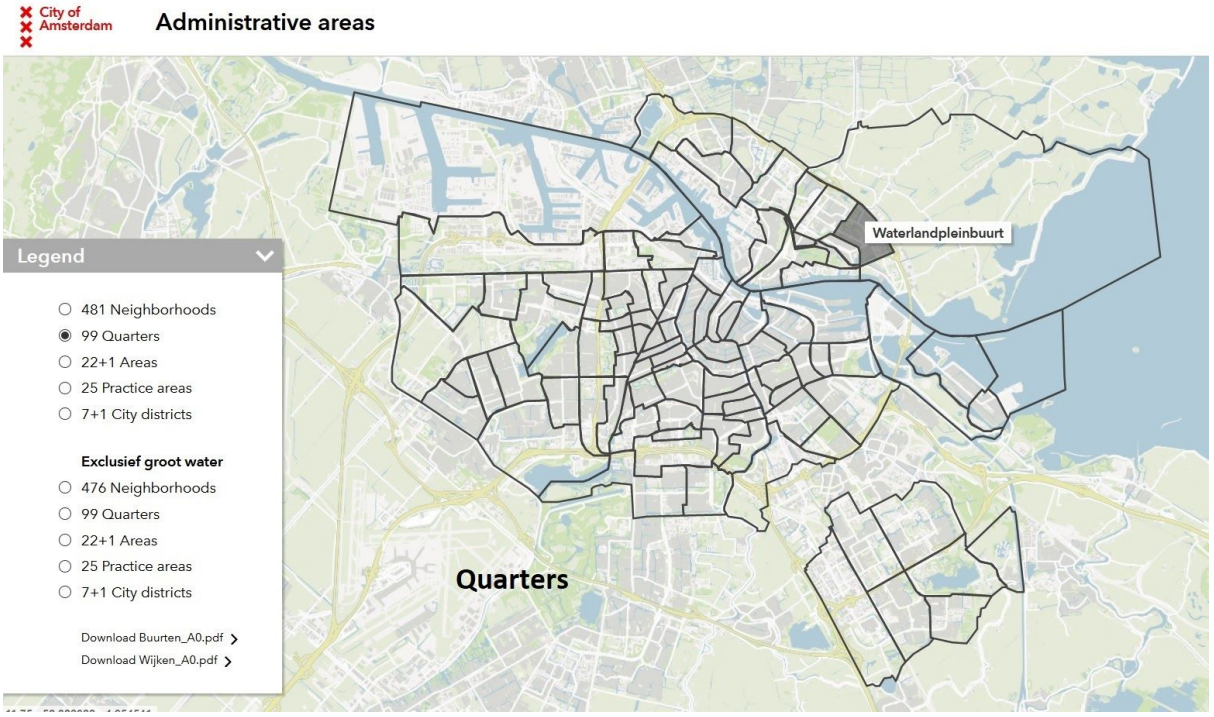
The data that we are going to use as part of this analysis has been taken from multiple locations. For instance, the data concerning the different neighbourhoods of Amsterdam was taken from the ‘clair city data portal’.

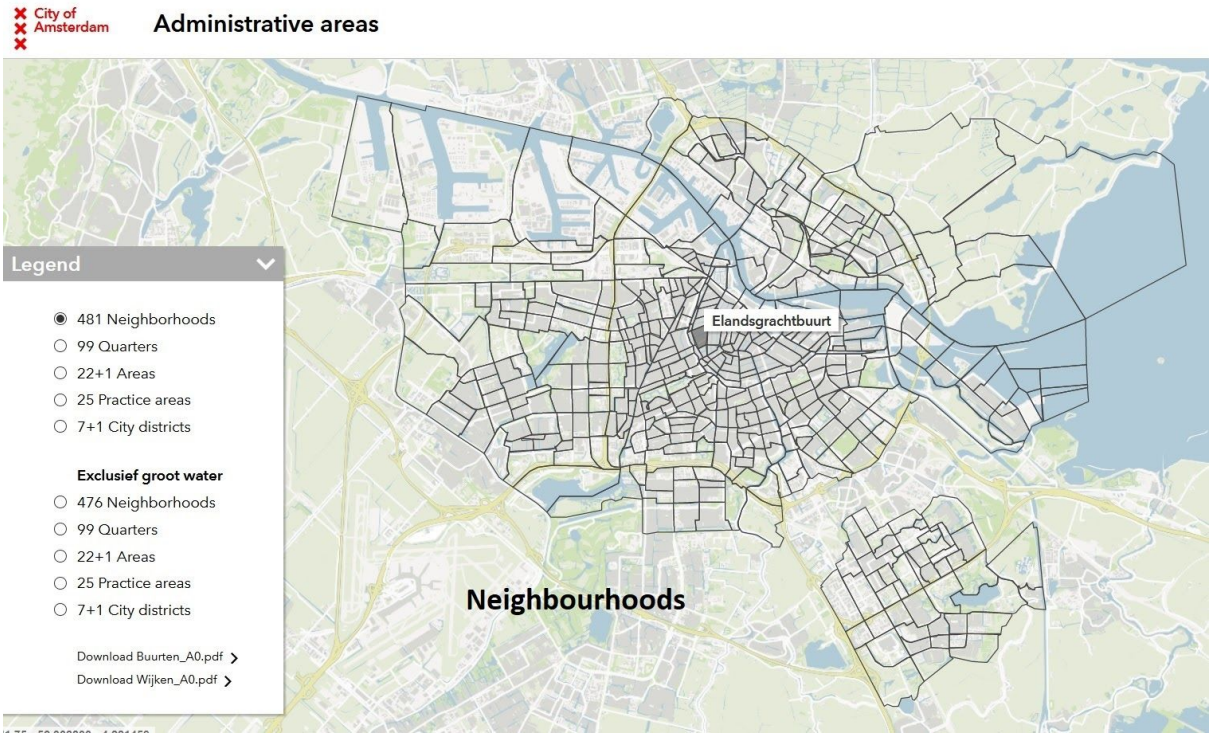
https://claircitydata.cbs.nl/organization/cbs

This dataset consisted of the region codes, type of region (quarter/neighbourhood), number of inhabitants, Latitude, Longitude etc. among other things. Since some of the columns and the data was in Dutch, I had to translate and research a bit about the neighbourhoods. I found a website of the Dutch government which gave me a pretty good understanding of the distribution of the landscape of Amsterdam.

https://maps.amsterdam.nl/gebiedsindeling/?LANG=en

Upon examination of the dataset and with the help of the above site, I found that *Amsterdam* has 99 quarters and 481 neighbourhoods in total. Basically a quarter (‘Wijk) is an area consisting of 2 or more neighbourhoods(‘Buurten’). So using this information I have extracted the data by filtering on quarters (‘Wijk’) as we would have a few neighbourhoods(‘Buurten’) to choose from at the end of our analysis.

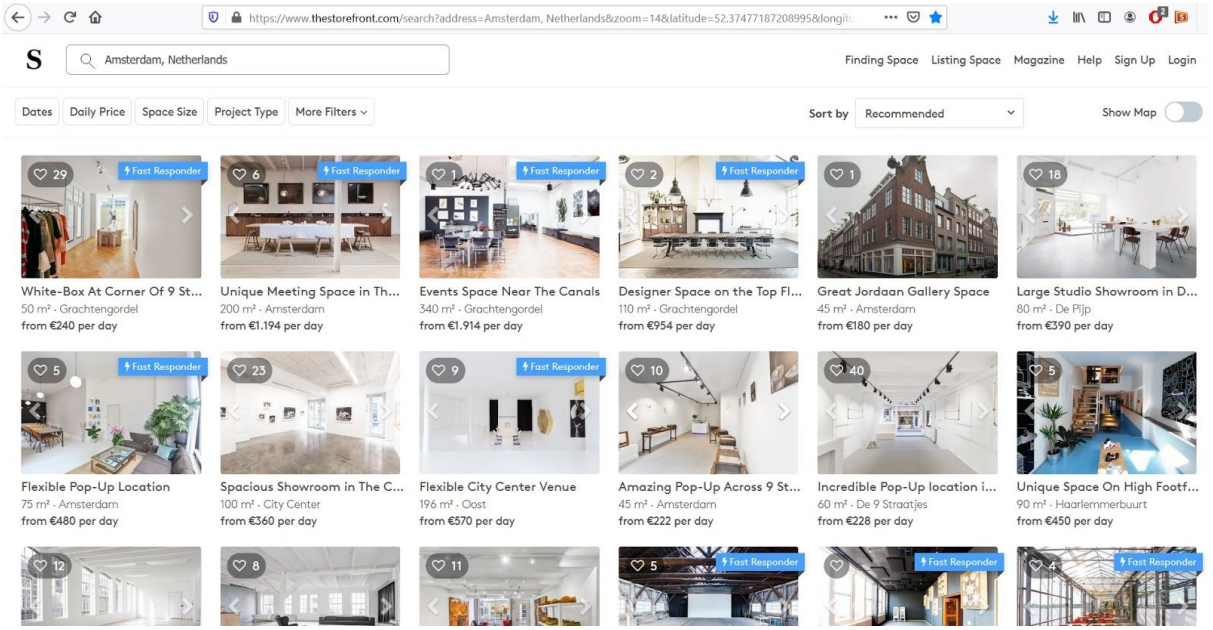




For the information regarding the venues with respect to the locations, we used the Foursquare API. We were able to get a list of venues for each neighbourhood corresponding to their latitude and longitude we extracted from the neighbourhoods dataset. For example,

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Burgwallen-Oude Zijde	52.372566	4.896943	Rosalia's Menagerie	52.371678	4.899174	Cocktail Bar
1	Burgwallen-Oude Zijde	52.372566	4.896943	De Koffieschenkerij	52.374043	4.898427	Coffee Shop
2	Burgwallen-Oude Zijde	52.372566	4.896943	Sofitel Legend The Grand Amsterdam	52.371093	4.895410	Hotel
3	Burgwallen-Oude Zijde	52.372566	4.896943	Wynand Fockink	52.372301	4.895253	Liquor Store
4	Burgwallen-Oude Zijde	52.372566	4.896943	Dam	52.372824	4.893702	Plaza

Apart from the above, we also needed data for commercial rented space for restaurants. Obtaining data for this part was a little bit tricky. However, I found a website which lets its users browse thousands of pop-up retail spaces, showrooms and event venues of a certain location.



Link : https://www.thestorefront.com/

I web scraped the data from the above site for all retail spaces in Amsterdam by adding a filter for 'Restaurants'. I extracted the location data (Latitude & Longitude) and the rental costs from the webpage into a pandas dataframe. Some of the spaces listed on the site had a 'Price on request' tag for which it was not possible to extract the rental cost data. For the purpose of our analysis all such listings had been removed from our dataset.

	id	Latitude	Longitude	Rent per day	Rent per month	Rent per week	residential
0	14993	52.364359	4.905604	480	14400	3360	Weesperbuurt
1	15624	52.369796	4.884745	240	4470	1320	Negen Straatjes
2	15851	52.364987	4.892312	1194	35820	8358	Grachtengordel
3	9647	52.368164	4.883889	1914	57420	13398	Negen Straatjes
4	9656	52.366548	4.892120	954	28620	6678	Grachtengordel
5	22615	52.372897	4.881112	180	2520	840	Jordaan
6	9680	52.353723	4.899191	390	11700	2730	de Pijp
7	16649	52.365795	4.877990	480	3600	1440	Helmersbuurt
8	14417	52.353952	4.910454	240	3000	1200	Weesperzijde
9	17362	52.363032	4.908441	570	17100	3990	Weesperbuurt
10	9651	52.371448	4.899676	360	10800	1680	Chinatown

3. Methodology :

To identify the segmentation of neighbourhoods with respect to the venues, we used K-means clustering on the venue location data which we obtained from the foursquare API. But before we applied K-means clustering on the data, we did an exploratory analysis on the venue data from the API.

3.1 Venue count :

With Amsterdam being a major tourist attraction bringing in people from all over the world, it would be helpful to have a look at the number of venues in each neighbourhood/quarter.

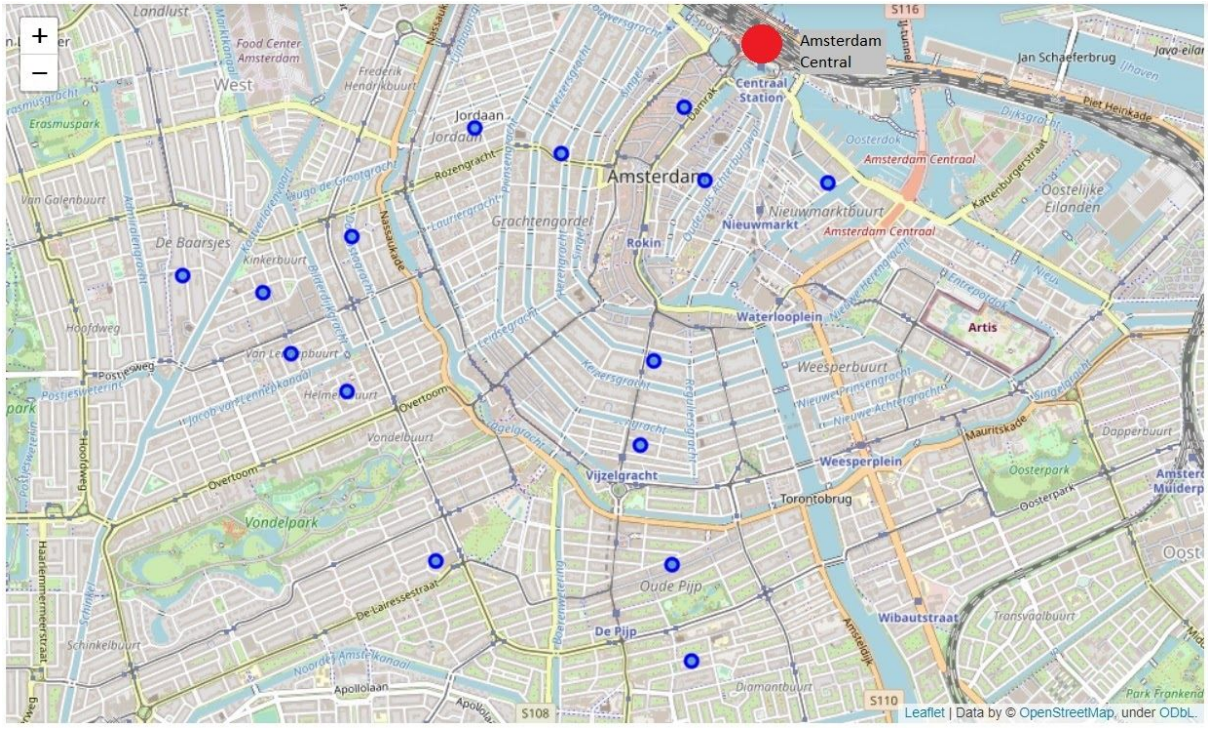
	Neighbourhood	count
0	Oude Pijp	100
1	Da Costabuurt	100
2	Grachtengordel-Zuid	100
3	Grachtengordel-West	100
4	Van Lennepbuurt	100
5	Museumkwartier	100
6	De Weteringschans	100
7	Jordaan	100
8	Burgwallen-Nieuwe Zijde	100
9	Burgwallen-Oude Zijde	100
10	Kinkerbuurt	100
11	Nieuwmarkt/Lastage	97
12	Nieuwe Pijp	95
13	Chassébuurt	78
14	Helmersbuurt	71
15	Vondelbuurt	70

As per the data extracted shown on the left , we have 11 quarters/neighbourhoods with a maximum of 100 number of venues. This is due to the reason that we put a limit of 100 on the number of venues returned by the Foursquare API for our analysis.

A quick browse on google would show you that all of these neighbourhoods with 100 venues are in close proximity to Amsterdam Central which is a major tourist hotspot.

Note: *Please note that the above is just a snapshot and not the entire list. Refer to the notebook in the github repository for complete list.*

A high volume of venues is something which would be preferable for our scenario as it means high footfall. Due to the centralised location and in close proximity to major tourist attractions, a restaurant in one of these neighbourhoods just might work in our favour. To visualize these neighbourhoods with a high venue count, we've used Folium to look at them on a map.



Note: *For the purpose of illustration, only a sample of neighbourhoods from the high venue count were shown on the map using Folium. Refer to the notebook in the github repository for the complete analysis.*

As you see above, all of the 'blue' pointers are in close proximity to 'Amsterdam Centraal' hardly 3-5 kms away and can be reached in appx. 10-15 mins via public transport.

3.2 Competition :

But it is not just the location which is a factor, it would also be helpful to find out how many venues out of the above are restaurants. Let us do a quick filter with respect to restaurants and sort them in descending order.

	Neighbourhood	Count of Restaurants
0	Kinkerbuurt	39
1	Van Lennepbuurt	38
2	Oude Pijp	36
3	Da Costabuurt	36
4	Nieuwe Pijp	32
5	Chassébuurt	27
6	Nieuwmarkt/Lastage	25
7	Helmersbuurt	24
8	Scheldebuurt	24
9	De Weteringschans	24
10	Grachtengordel-Zuid	23
11	Indische Buurt West	22
12	Jordaan	20
13	Museumkwartier	20
14	Burgwallen-Oude Zijde	19
15	Weesperzijde	17
16	Grachtengordel-West	16
17	Burgwallen-Nieuwe Zijde	15

Even after filtering 'restaurants' from the venues of all the neighbourhoods, you'll notice that the most of the neighbourhoods at the top of this list match with the previous list we showed.

There are still quite a sufficient amount of eateries in these neighbourhoods which are in close proximity to the tourist hotspots.

There are quite a sufficient amount of restaurants in the neighbourhoods. After all Amsterdam is quite a tourist attraction. But since we're looking to open an *Italian* themed restaurant, let us check how many are present in the data that we've extracted. It does help to know the competition!

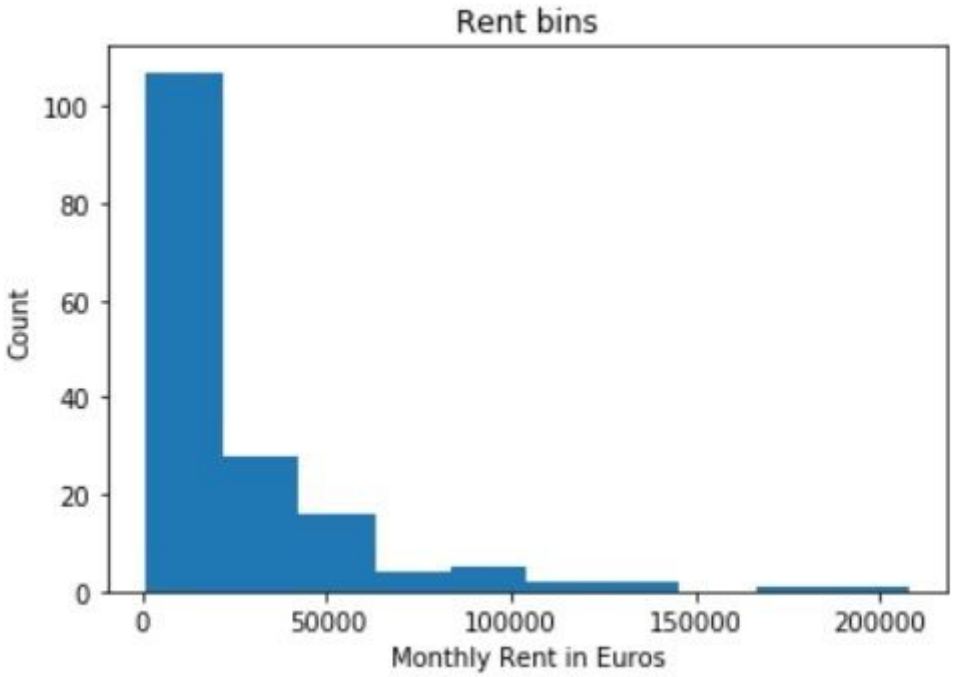
	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
85	Burgwallen-Oude Zijde	52.372566	4.896943	La Zoccola del Pacioccone	52.375297	4.893965	Italian Restaurant
107	Burgwallen-Nieuwe Zijde	52.375703	4.895518	La Zoccola del Pacioccone	52.375297	4.893965	Italian Restaurant
183	Burgwallen-Nieuwe Zijde	52.375703	4.895518	Pastabar	52.376518	4.897432	Italian Restaurant
189	Burgwallen-Nieuwe Zijde	52.375703	4.895518	Ava Cyrill İtaliaans Restaurant	52.377954	4.895812	Italian Restaurant
192	Burgwallen-Nieuwe Zijde	52.375703	4.895518	Il Pacioccone	52.376499	4.895622	Italian Restaurant

There are quite a few Italian restaurants spread across Amsterdam as you can see in the images, so it would be wise to see the results of our cluster analysis before deciding the optimal neighbourhood to open our restaurant in. A better idea would be to filter the count per neighbourhood and decide the neighbourhood with minimal count along with the cluster analysis.

	Neighbourhood	Venue Count	Italian Restaurant count
0	Grachtengordel-West	100	2
1	Grachtengordel-Zuid	100	7
2	Jordaan	100	2
3	Da Costabuurt	100	5
4	Museumkwartier	100	2
5	Burgwallen-Oude Zijde	100	1
6	Burgwallen-Nieuwe Zijde	100	4
7	De Weteringschans	100	4
8	Oude Pijp	100	2
9	Van Lennepbuurt	100	4
10	Kinkerbuurt	100	4
11	Nieuwe Pijp	94	5
12	Nieuwmarkt/Lastage	90	4
13	Chassébuurt	79	3
14	Helmersbuurt	71	3

3.3 The Rent Factor:

Another factor which is crucial while opening a business that we discussed earlier was the rental costs of the establishment. As we scraped the data from 'TheStorefront.com' into a pandas dataframe, it would be helpful to visualize the distribution of rental costs.



From the above histogram we can see that the majority of our data is left-skewed i.e most of the establishments 'Rent per month' is below 20,000€ which is still a lot. So for our analysis, we filtered the list down to properties with rent under 5,000€ per month.

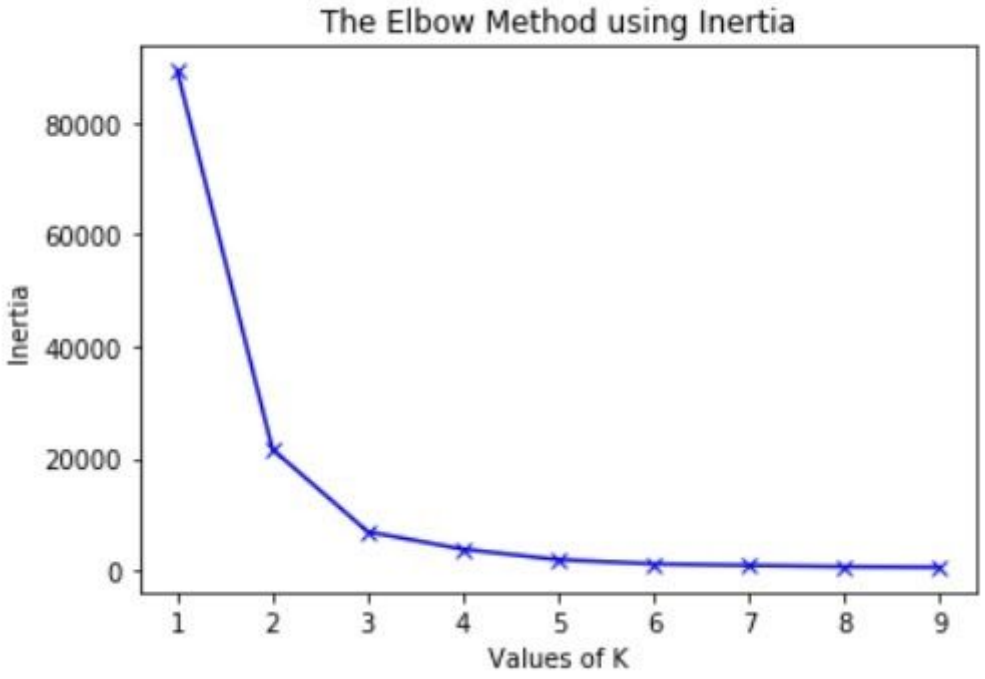
Note: *Only a partial snapshot shown, complete list in notebook on github repository*

	id	Latitude	Longitude	Rent per day	Rent per month	Rent per week	residential
52	21466	52.373548	4.888542	30	900	210	Negen Straatjes
33	9687	52.370135	4.890675	78	1428	540	Centrum
77	9686	52.369711	4.889050	51	1428	357	Burgwallen-Nieuwe Zijde
48	9657	52.373744	4.880662	50	1500	350	Jordaan
51	22116	52.379728	4.892223	86	1800	600	Haarlemmerbuurt
110	13805	52.378777	4.899093	60	1800	420	Centrum
46	9567	52.371642	4.880424	300	2400	780	Jordaan
5	22615	52.372897	4.881112	180	2520	840	Jordaan
136	9639	52.367272	4.890582	84	2520	588	Centrum
135	18138	52.368749	4.888586	112	2580	780	Centrum
79	9645	52.372031	4.890857	88	2640	616	Centrum
57	16296	52.373031	4.904094	252	2640	870	Nieuwmarktbuurt
37	14681	52.354897	4.900224	90	2700	630	de Pijp

3.4 Clustering:

Now, let us apply K-means clustering on the neighbourhoods/quarters of Amsterdam to see the segmentation. After obtaining the data from the Foursquare API and summing the total number of venues for each category we applied K-means to the data set. For the purpose of finding the best neighbourhood with the highest footfall we also added the venue count to our analysis since we wanted it to be the main driving force behind the clustering.

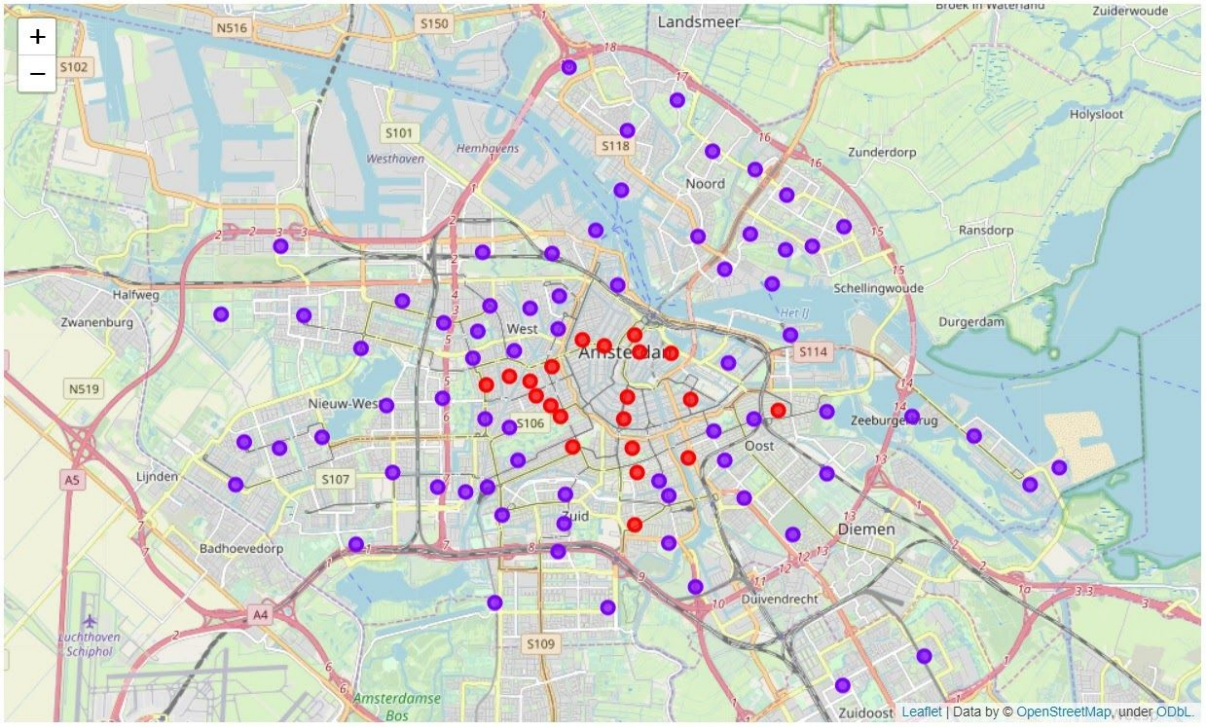
To ensure we had the optimal number of clusters for K-means and did not miss any hidden observations, we used the 'Elbow method' as well as 'Avg. Silhouette score' for each data point.



Average Silhouette Score:

```
For no of clusters = 2  The average silhouette_score is : 0.7034679488173275
For no of clusters = 3  The average silhouette_score is : 0.6701166121669848
For no of clusters = 4  The average silhouette_score is : 0.6382717865777569
For no of clusters = 5  The average silhouette_score is : 0.6599014857542905
For no of clusters = 6  The average silhouette_score is : 0.6692516146009787
For no of clusters = 7  The average silhouette_score is : 0.6114119423979929
For no of clusters = 8  The average silhouette_score is : 0.6020677608611836
For no of clusters = 9  The average silhouette_score is : 0.5950956924116599
For no of clusters = 10  The average silhouette_score is : 0.6203731789194354
For no of clusters = 11  The average silhouette_score is : 0.6126705120734399
For no of clusters = 12  The average silhouette_score is : 0.6073634599804639
```

Using both the above methods we clearly have an optimal number of clusters to be used for K-means which is 2.



Looking at the above map we see that ‘Cluster 0’ is localized to the centre of Amsterdam with all the tourist locations and the high venue counts and ‘Cluster 1’ is more to the exterior from the city centre. So, we chose Cluster 1 as that would give us a higher footfall and a better location. Upon further examination of the venues in this cluster we notice there are a lot of Bars, Pubs, Marijuana Dispensary, Restaurants etc in this cluster as it is closer to the city centre which has pretty famous ‘Nightlife’.

	Neighbourhood	City	Region code	Number of inhabitants	Longitude	Latitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Burgwallen-Oude Zijde	Amsterdam	WK036300	4280	4.896943	52.372566	0	Bar	Hotel	Coffee Shop	Museum	Marijuana Dispensary
1	Burgwallen-Nieuwe Zijde	Amsterdam	WK036301	3960	4.895518	52.375703	0	Bar	Coffee Shop	Italian Restaurant	Dessert Shop	Bakery
2	Grachtengordel-West	Amsterdam	WK036302	6475	4.886836	52.373709	0	Bar	Hotel	Café	French Restaurant	Marijuana Dispensary
3	Grachtengordel-Zuid	Amsterdam	WK036303	5440	4.893349	52.364821	0	Italian Restaurant	Marijuana Dispensary	Hotel	Movie Theater	Coffee Shop
4	Nieuwmarkt/Lastage	Amsterdam	WK036304	9605	4.905641	52.372436	0	Bar	Hotel	Coffee Shop	Italian Restaurant	Chinese Restaurant
5	Haarlemmerbuurt	Amsterdam	WK036305	9305	4.890582	52.384171	1	Bar	Marijuana Dispensary	Restaurant	Gastropub	Organic Grocery

4. Results :

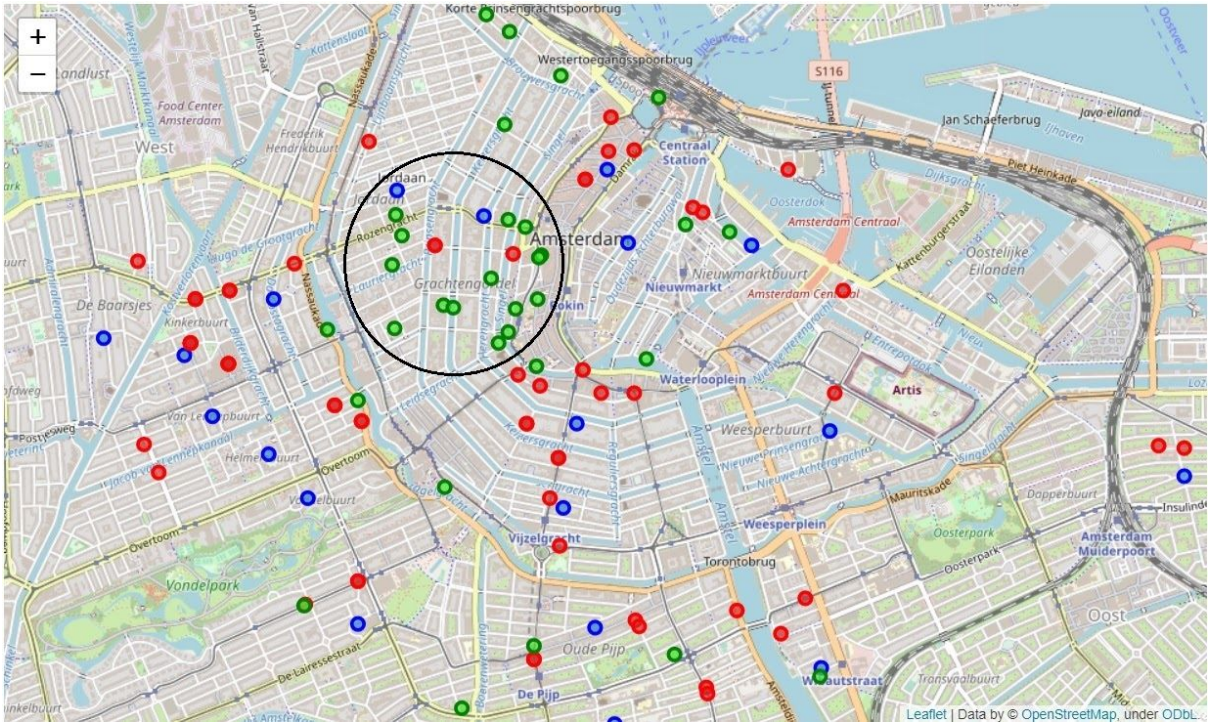
After running the K-means Cluster we basically obtained 2 clusters corresponding to their categories as follows:

Cluster Labels	Category
Cluster 0	High venue count, high footfall, close to city centre
Cluster 1	Low venue count, lower footfall, residential/office areas

So, we filtered the quarters/neighbourhoods in Cluster 0 and we also filtered the ‘Italian Restaurants’ which are part of these neighbourhoods (Basically our competition). Now, if you remember we also had a dataframe with venue count as well as the count of Italian restaurants, so we would choose a neighbourhood filtered with our cluster with a minimal count.

	Neighbourhood	Venue Count	count	Longitude	Latitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Burgwallen-Oude Zijde	100	1	4.896943	52.372566	Bar	Hotel	Coffee Shop	Museum	Marijuana Dispensary
1	Burgwallen-Nieuwe Zijde	100	4	4.895518	52.375703	Bar	Coffee Shop	Italian Restaurant	Dessert Shop	Bakery
2	Grachtengordel-West	100	2	4.886836	52.373709	Bar	Hotel	Café	French Restaurant	Marijuana Dispensary
3	Grachtengordel-Zuid	100	7	4.893349	52.364821	Italian Restaurant	Marijuana Dispensary	Hotel	Movie Theater	Coffee Shop
4	Nieuwmarkt/Lastage	90	4	4.905641	52.372436	Bar	Hotel	Coffee Shop	Italian Restaurant	Chinese Restaurant
5	Jordaan	100	2	4.880786	52.374805	Bar	Café	Hotel	Coffee Shop	Yoga Studio
6	De Weteringschans	100	4	4.892427	52.361170	Bar	Coffee Shop	Café	Italian Restaurant	French Restaurant
7	Weesperbuurt/Plantage	66	1	4.911103	52.364499	Zoo Exhibit	Hotel	Bar	Restaurant	Café
8	Da Costabuurt	100	5	4.872120	52.370131	Bar	Coffee Shop	Italian Restaurant	Yoga Studio	Dutch Restaurant

After filtering, we overlapped all these locations, i.e the neighbourhoods we chose as part of cluster analysis, the ‘Italian restaurants’ which are our competition and the properties with monthly rental cost below 5.000 euros on a map using Folium. This would help us in making our final decision to choose the optimum neighbourhood.



In the above map,
Red highlights our competition, i.e the other ‘Italian’ restaurants.
Blue highlights the neighbourhoods we chose from the Cluster analysis.
Green highlights the properties with monthly rentals below 5.000 Euros.

At a closer look we find 3 neighbourhoods namely ‘Grachtengordel-West’, ‘Jordaan’ and ‘Burgwallen-Oude-Zijde’ which have minimal competition and a lot of properties to choose from.

5. Discussion :

Apart from the above analysis, we notice that the second cluster has a lot of residential population and a less number of venues. A fact which can be substantiated from the dataset containing the neighbourhood information along with the ‘number of inhabitants’ and also the Foursquare API. If you notice our analysis has mainly taken into account the fact that Amsterdam is a major tourist location with a high footfall in the city centre due to the many tourist locations and nightlife venues.

Suppose we were to target the residential population, then we could analyze data with regards to neighbourhoods having a lot of Italian immigrants and/or expatriates. Another business idea would be to open a take-away service at those locations, cross-referencing them with the property data to find low-cost rental establishments. Opening a restaurant near to such a location might transform to a lot of revenue. Most parts of our analysis might hold good for a different business venture if we could find the right market for our product.

Also, if you notice we’ve put a cap on the monthly rental costs to below 5.000 Euros for our analysis. For people with a lower budget or perhaps an extravagant budget, that filter could be adjusted and a whole lot of other properties might have an altogether different outcome.

6. Conclusion :

Based on the above detailed analysis, ‘Grachtengordel-West’ and ‘Jordaan’ would be the best neighbourhood choices considering all the factors : high footfall, close to the city centre and tourist locations, low to medium rent options and the competition. Another neighbourhood that can be considered is ‘Burgwallen-Oude-Zijde’.
‘Oude Pijp, ‘Weesperbuurt/Plantage’ & ‘Haarlemmerbuurt’ are some of the other neighbourhoods which could have been considered if they had a lot of commercial property options.