FINDING THE BEST NEIGHBORHOOD TO MOVE IN

Murad Hüdavendigar BOZIK

July 21, 2019

1. Introduction

1.1 Scenario

You are a great data scientist with a love of beating your personal bests and currently live in Istanbul, Turkey. You receive a job offer from a significant company in Amsterdam, Netherlands. If you decide to accept the offer which is indisputable well opportunity for your career, you have to move to Amsterdam city. In order to show your best performance, you want to keep adaptation period as short as possible.

1.2 Problem

When you move in Amsterdam the biggest problem for you will be preparing food. Because in your home town you can arbitrarily go to a restaurant and you can find the meals that you liked. Since you work hard in office most probably you will stay so long at beginning. You may not find time to prepare food yourself. For that reason, you're expecting to move in a neighbourhood which is similiar to your home town in terms of a venues in vicinity. You can find tastes you like and it would not effect your performance.

This project aims to find an optimal location for you with turkish restaurants in vicinity and similar location to your home town in terms of nearby venues.

2. Data acquisition and cleaning

2.1 Data sources

We need two datasets. Each dataset should include at least following features;

- Districts
- Neighbourhoods
- Geographical coordinates (latitude, longitude)

✔ In the website of amsterdam municipality there is neighbourhoods open dataset, we will use it for our analysis. You can reach this data set with this link: City of Amsterdam neighbourhoods dataset. Unfortunately, there is no ready-to-use dataset for the city of Istanbul. We will scrape the required fetures from this website link: List of neighbourhoods of Istanbul.

Following data sources will be needed to extract/generate the required information:

✔ Geographical coordinates of neighbourhoods in Istanbul will be obtained using **Google Maps API** geocoding

✔ Number of venues and their type and location in every neighborhood will be obtained using **Foursquare API**

2.2 Data Cleaning and feature selection

2.2.1 Amsterdam Dataset

In amsterdam dataset there were lots redundant features. Simply we selected required features;

✗ OBJECTNUMMER : Represents numbers as index

✗ Buurt_code : Unique Neighbourhoods code

✔ Buurt : Neighbourhoods name

✗ Buurtcombinatie_code : Another Neighbourhoods code

✔ Stadsdeel_code : Districts code

✗ Opp_m2 : Surface area in m2 unit

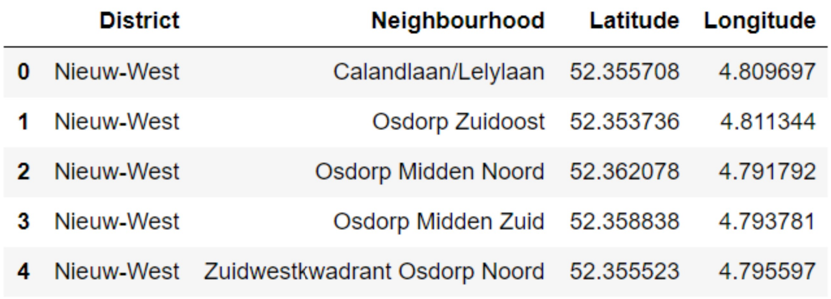
✗ WKT_LNG_LAT : Polygon longitude contours

✗ WKT_LAT_LNG : Polygon latitude contours

✔ LNG : Longitude

✔ LAT : Latitude

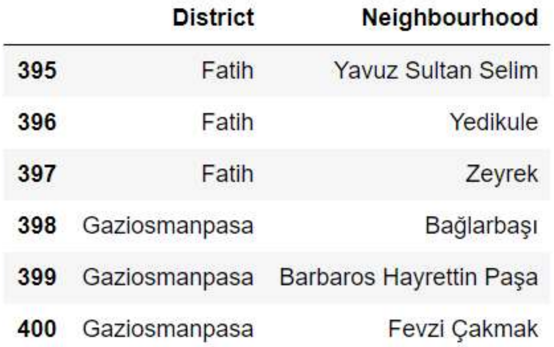
After selection and renamed features;



	District	Neighbourhood	Latitude	Longitude
0	Nieuw-West	Calandlaan/Lelylaan	52.355708	4.809697
1	Nieuw-West	Osdorp Zuidoost	52.353736	4.811344
2	Nieuw-West	Osdorp Midden Noord	52.362078	4.791792
3	Nieuw-West	Osdorp Midden Zuid	52.358838	4.793781
4	Nieuw-West	Zuidwestkwadrant Osdorp Noord	52.355523	4.795597

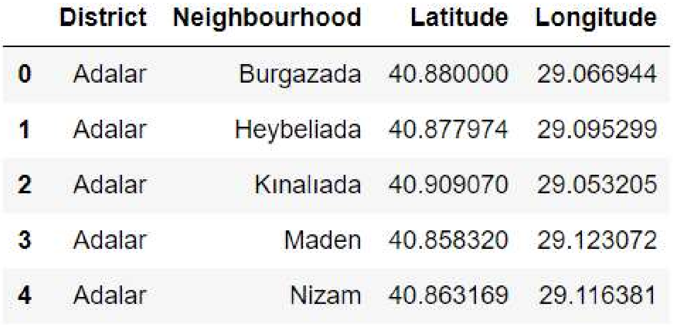
2.2.2 Istanbul Dataset

After scraping data from the link;



	District	Neighbourhood
395	Fatih	Yavuz Sultan Selim
396	Fatih	Yedikule
397	Fatih	Zeyrek
398	Gaziosmanpasa	Bağlarbaşı
399	Gaziosmanpasa	Barbaros Hayrettin Paşa
400	Gaziosmanpasa	Fevzi Çakmak

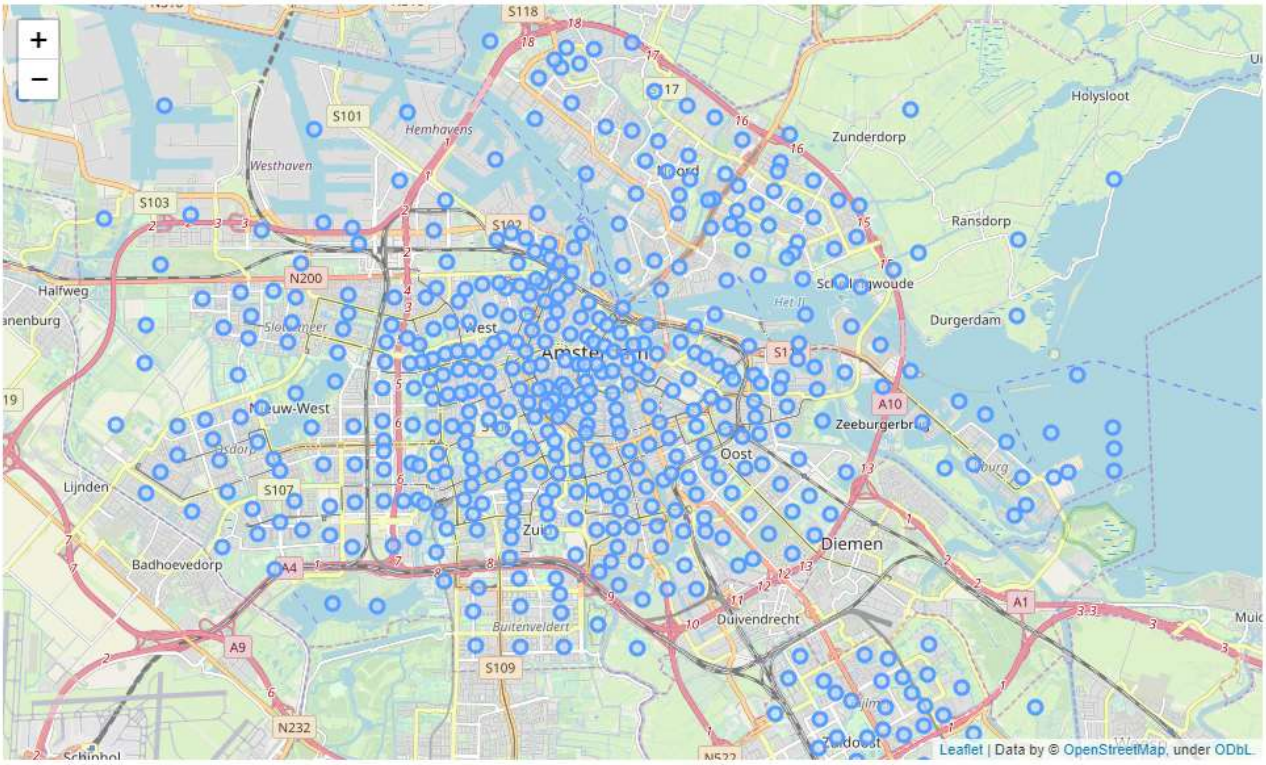
We used Google API geocoding for getting geographical coordinates of each neighbourhood. We dropped all rows with missing location data. After all cleaning process;



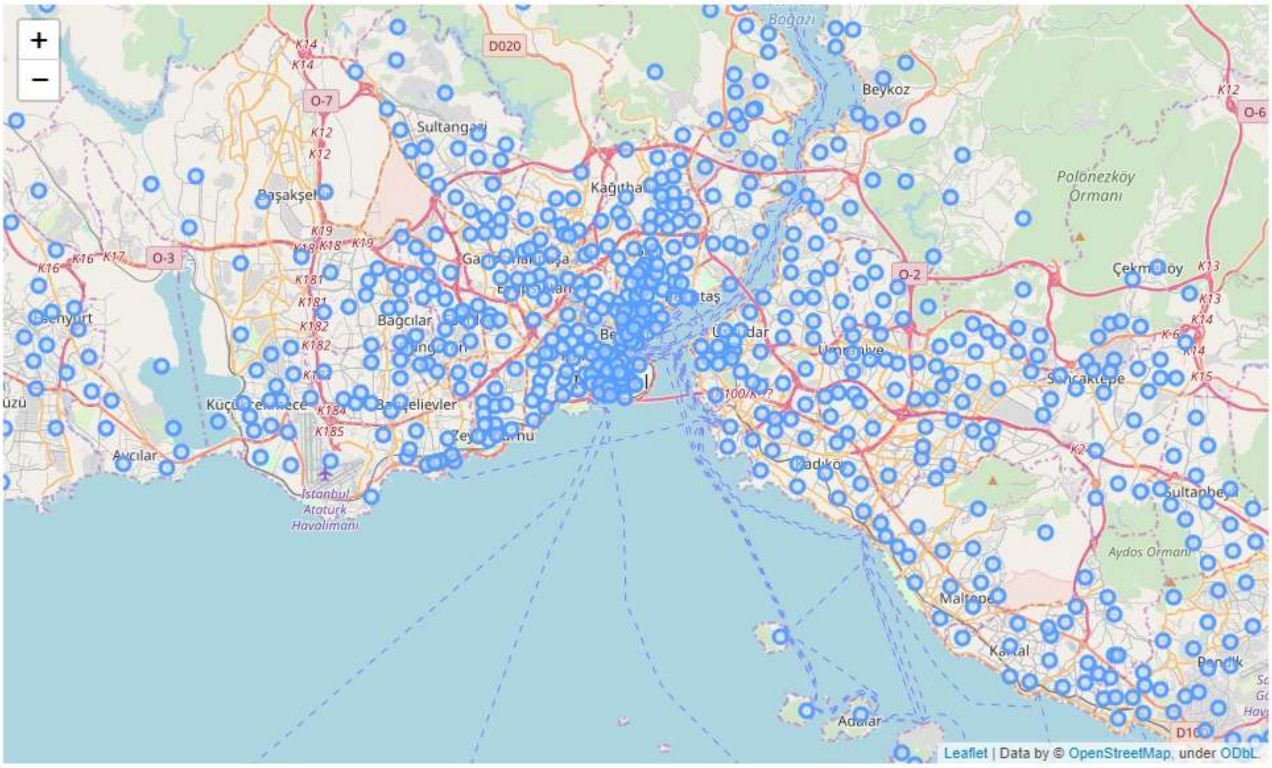
	District	Neighbourhood	Latitude	Longitude
0	Adalar	Burgazada	40.880000	29.066944
1	Adalar	Heybeliada	40.877974	29.095299
2	Adalar	Kınalıada	40.909070	29.053205
3	Adalar	Maden	40.858320	29.123072
4	Adalar	Nizam	40.863169	29.116381

2.2.3 Visualization of Datasets

Amsterdam

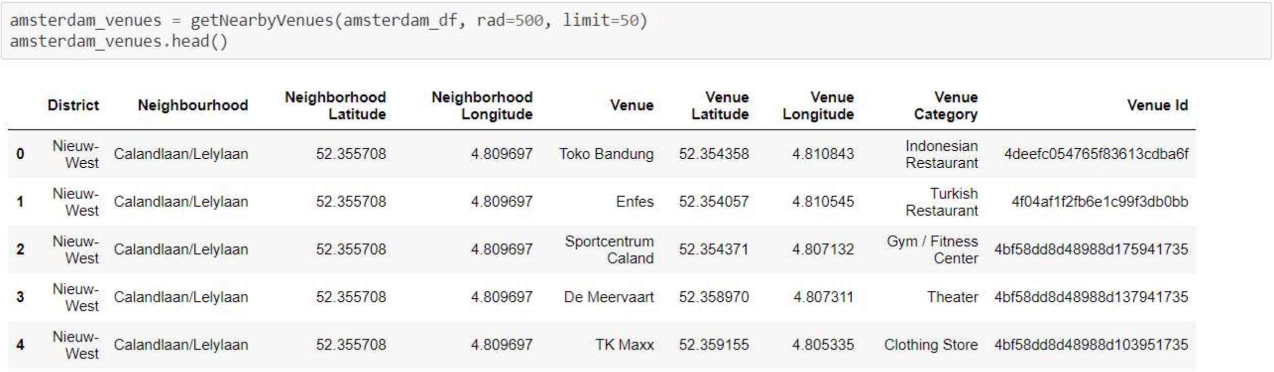


Istanbul



2.2.3. Foursquare

We used Foursquare API in order to get nearby venues information for each datasets. After getting venue info:



```
amsterdam_venues = getNearbyVenues(amsterdam_df, rad=500, limit=50)
amsterdam_venues.head()
```

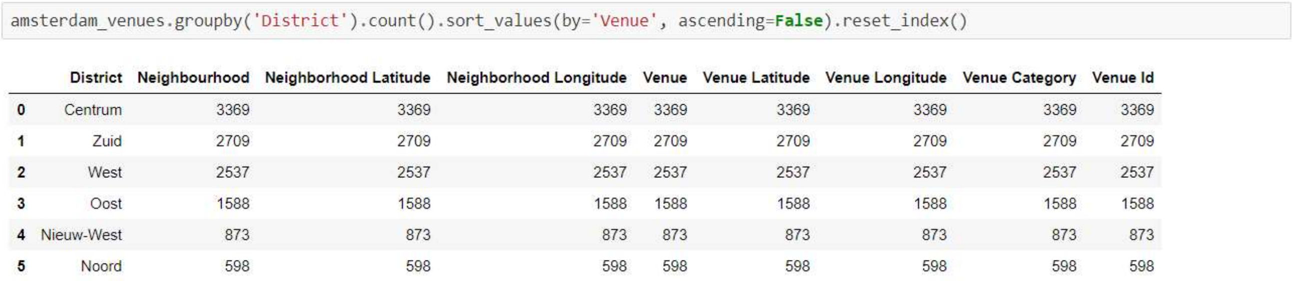
	District	Neighbourhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue Id
0	Nieuw-West	Calandlaan/Lelylaan	52.355708	4.809697	Toko Bandung	52.354358	4.810843	Indonesian Restaurant	4deefc054765f83613cdba6f
1	Nieuw-West	Calandlaan/Lelylaan	52.355708	4.809697	Enfes	52.354057	4.810545	Turkish Restaurant	4f04af1f2fb6e1c99f3db0bb
2	Nieuw-West	Calandlaan/Lelylaan	52.355708	4.809697	Sportcentrum Caland	52.354371	4.807132	Gym / Fitness Center	4bf58dd8d48988d175941735
3	Nieuw-West	Calandlaan/Lelylaan	52.355708	4.809697	De Meervaart	52.358970	4.807311	Theater	4bf58dd8d48988d137941735
4	Nieuw-West	Calandlaan/Lelylaan	52.355708	4.809697	TK Maxx	52.359155	4.805335	Clothing Store	4bf58dd8d48988d103951735

3. Explanatory Analysis

3.1. Amsterdam

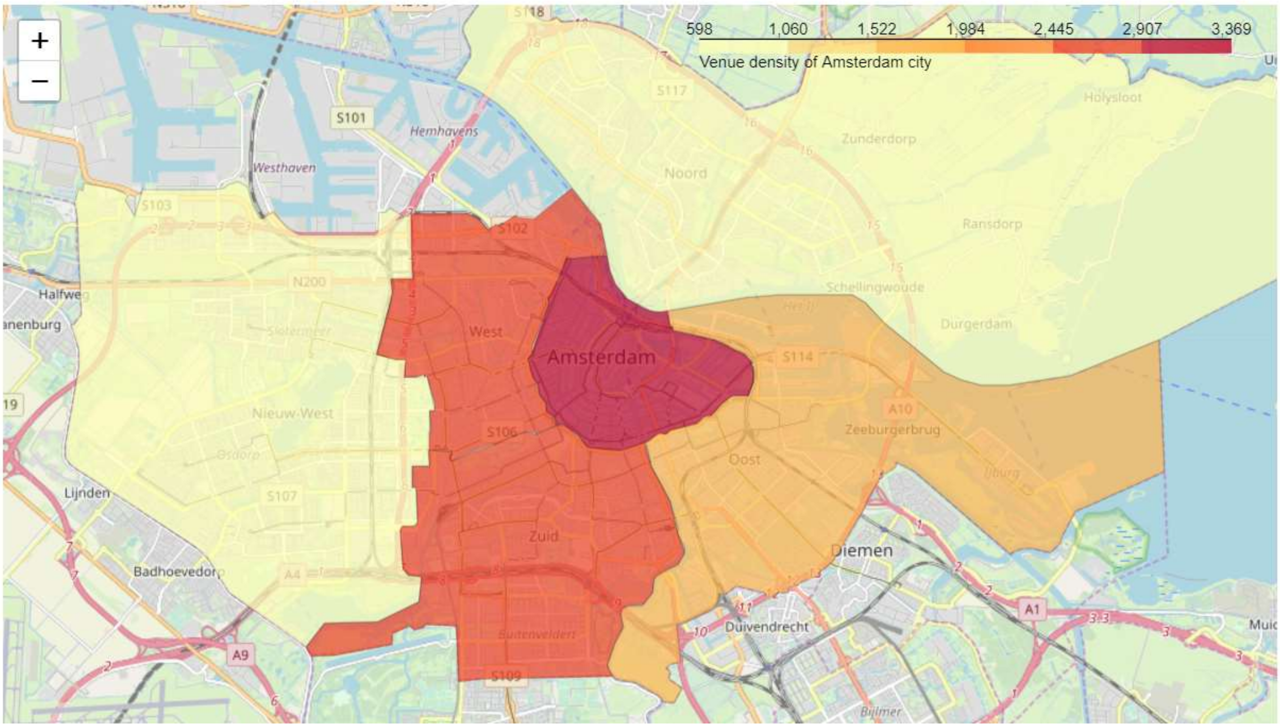
Regarding basic explanatory analysis we come up with these questions;

How many venues in Amsterdam city each district has? This question also shows the density distribution of venues overall city of Amsterdam.



```
amsterdam_venues.groupby('District').count().sort_values(by='Venue', ascending=False).reset_index()
```

	District	Neighbourhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue Id
0	Centrum	3369	3369	3369	3369	3369	3369	3369	3369
1	Zuid	2709	2709	2709	2709	2709	2709	2709	2709
2	West	2537	2537	2537	2537	2537	2537	2537	2537
3	Oost	1588	1588	1588	1588	1588	1588	1588	1588
4	Nieuw-West	873	873	873	873	873	873	873	873
5	Noord	598	598	598	598	598	598	598	598



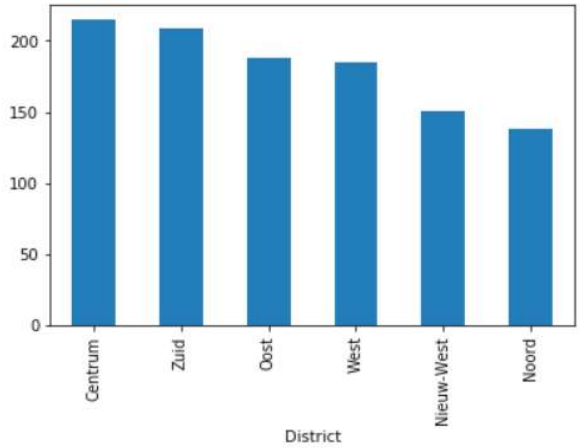
As we see on the map, most dense district is the city center as we expected. However, how much different their categories?

```
# Total unique categories number
amsterdam_venues['Venue Category'].nunique()
```

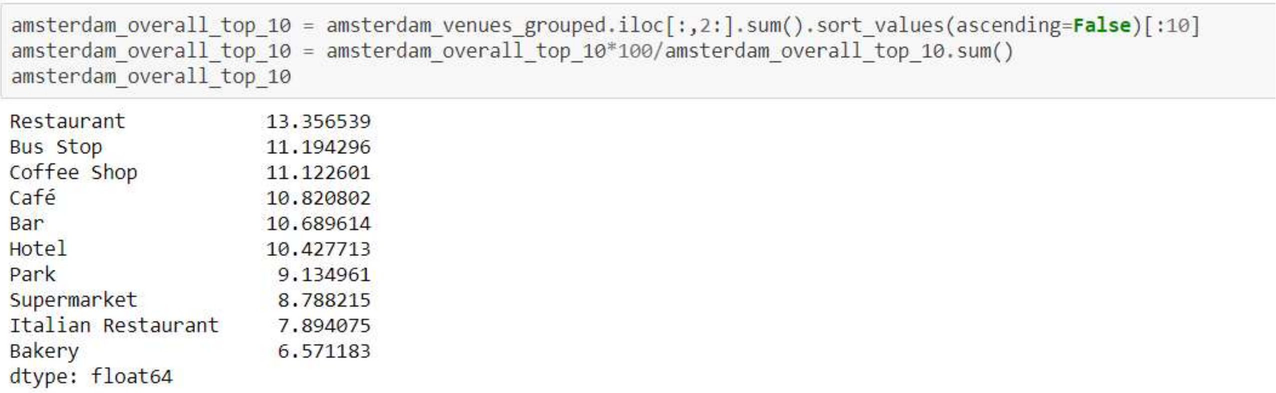
```
342
```

```
# Number of unique categories per district
y = amsterdam_venues.groupby('District')['Venue Category'].nunique().sort_values(ascending=False)
y.plot(kind='bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1f93ce451d0>
```



Next question is “What is most common top-10 venue category overall Amsterdam?”

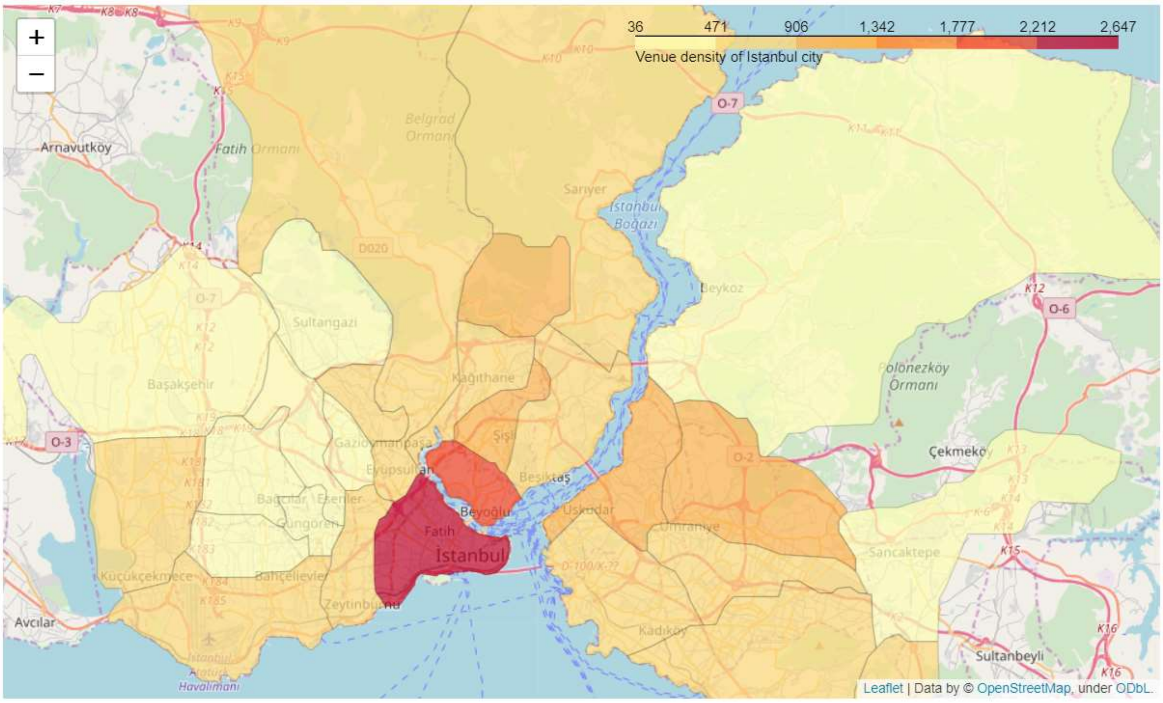


```
amsterdam_overall_top_10 = amsterdam_venues_grouped.iloc[:,2:].sum().sort_values(ascending=False)[:10]
amsterdam_overall_top_10 = amsterdam_overall_top_10*100/amsterdam_overall_top_10.sum()
amsterdam_overall_top_10
```

```
Restaurant           13.356539
Bus Stop             11.194296
Coffee Shop          11.122601
Café                 10.820802
Bar                  10.689614
Hotel                10.427713
Park                  9.134961
Supermarket           8.788215
Italian Restaurant    7.894075
Bakery                6.571183
dtype: float64
```

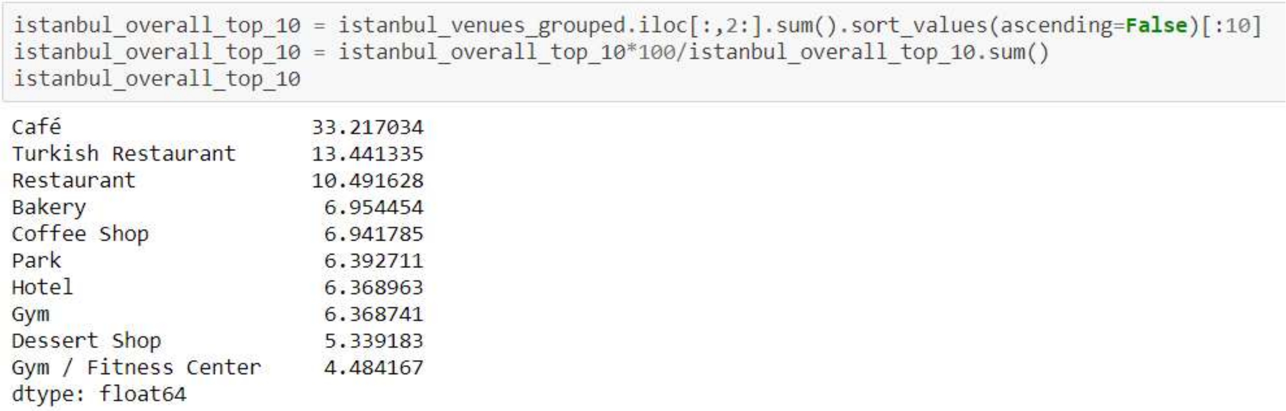
3.2. Istanbul City

Same question for Istanbul : What is the distribution of venue density?



Fatih district is the most dense district. And this district is also historical region of Istanbul.

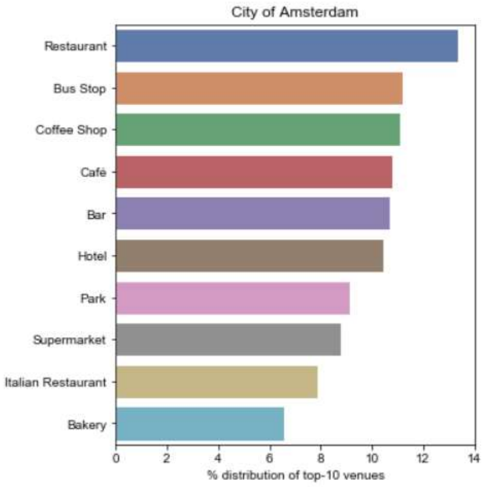
What is top-10 venue category for Istanbul city?

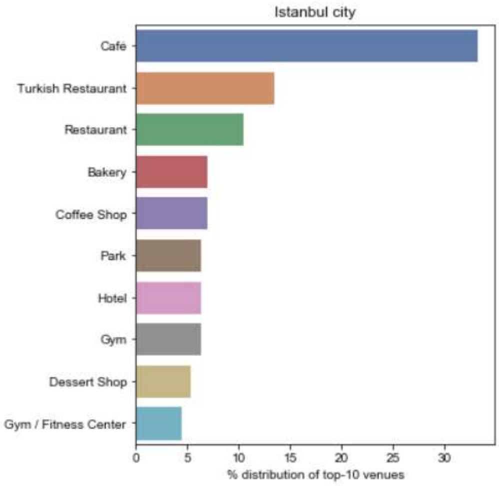


```
istanbul_overall_top_10 = istanbul_venues_grouped.iloc[:,2:].sum().sort_values(ascending=False)[:10]
istanbul_overall_top_10 = istanbul_overall_top_10*100/istanbul_overall_top_10.sum()
istanbul_overall_top_10
```

```
Café                    33.217034
Turkish Restaurant      13.441335
Restaurant              10.491628
Bakery                   6.954454
Coffee Shop              6.941785
Park                     6.392711
Hotel                    6.368963
Gym                      6.368741
Dessert Shop             5.339183
Gym / Fitness Center     4.484167
dtype: float64
```

We can compare two cities graphically.





We can see in the right graph in Turkish community, percentage of Cafe is the highest. It's rate is more that twice of those following category. On the other hand, In Dutch community, there is a balance between categories of top 10 venues.

4. Methodology

In order to find the best solution to our business problem, we will focus on how similar the neighbourhoods are based on venues information in vicinity. We got top-30 venue category for each neighbourhood in each city within 500 meters radius using Foursquare API. After we found the similar neighbourhoods to your town, we looked for neighbourhoods that involves Turkish venues. Basically we offer these venues because they are not only similar to your home town but also they have Turkish venues contain the tastes you wish.

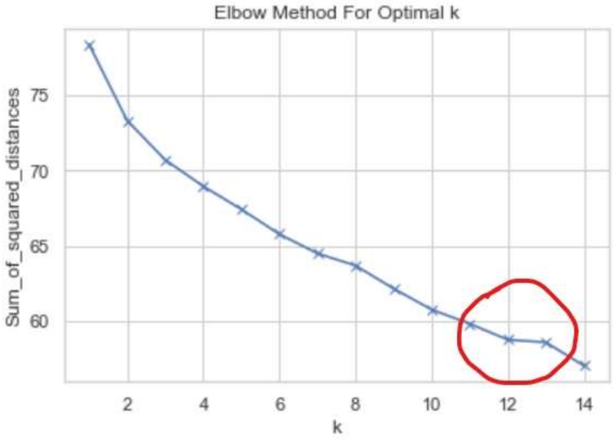
5. Modeling

We have no labeled data, therefore the problem of this project is an unsupervised problem. We simply apply a segmentation on neighbourhoods. We used K-Means algorithm to identify similarity of neighborhoods by clustering them. In k-means algorithm we need to identify the best cluster number. To identify that we used elbow method.

5.1. Identifying the best cluster number for K-Means

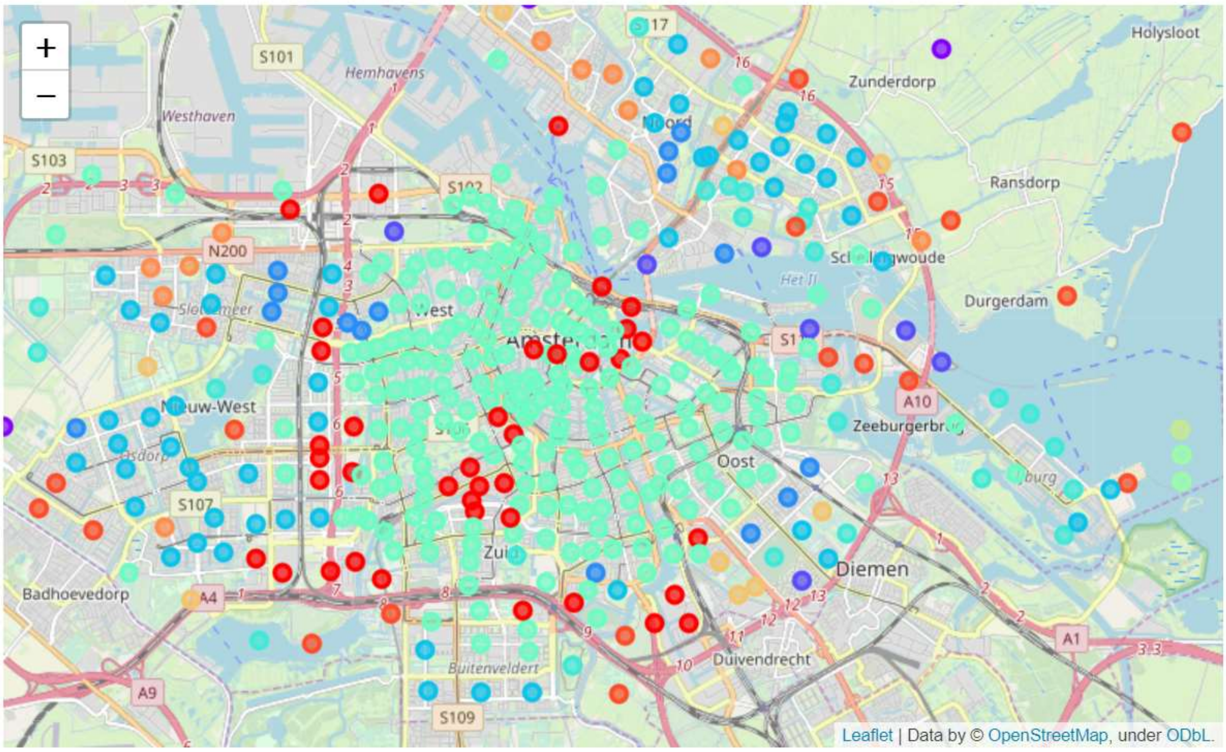
Elbow method we used is an observational analysis. We define a range for k number and for every k value we fit our model and calculate sum of squared distances. We know that if we increase cluster number sum of squared distances (ssd) always decrease. But the point is we should identify the most sognificant bend in graphic of ssd. We choose 15 as our k range.

From observation we can say that the most significant bend is at 12. So the best cluster number we should choose is 12.

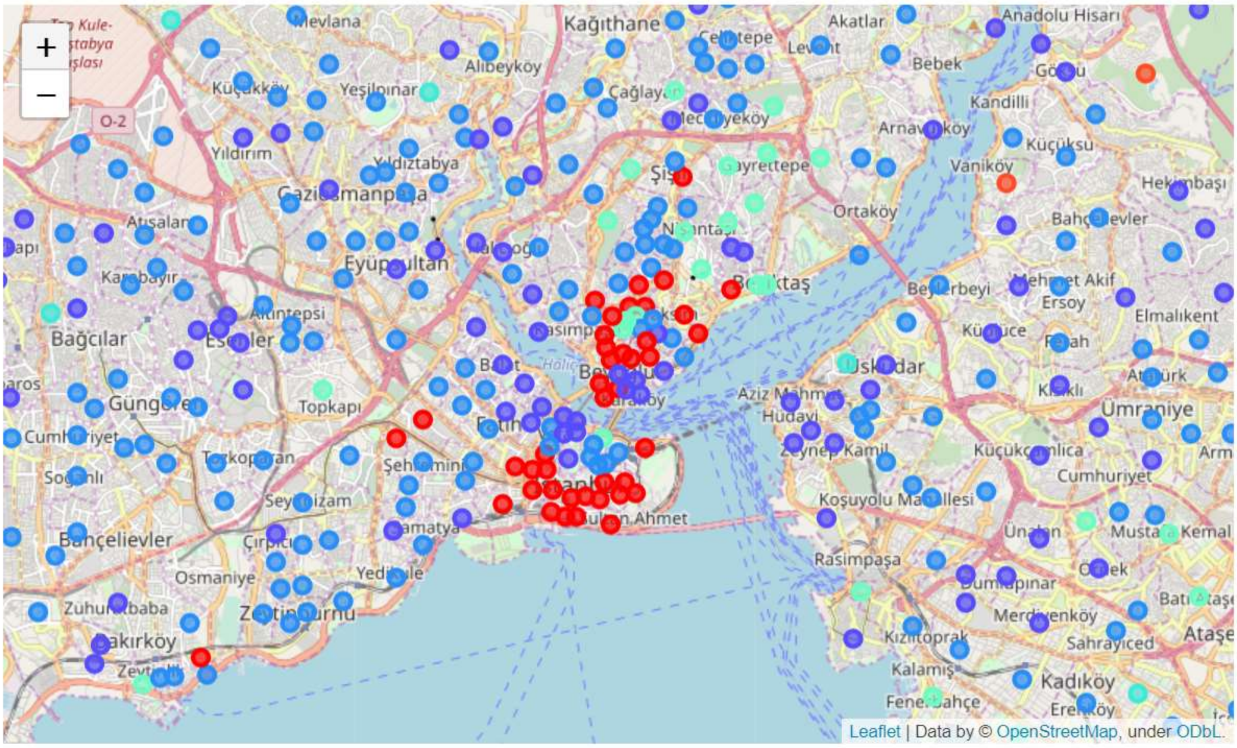


5.2 Results of clustering

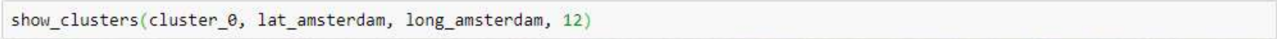
```
show_clusters(amsterdam_merged, lat_amsterdam, long_amsterdam, 12)
```



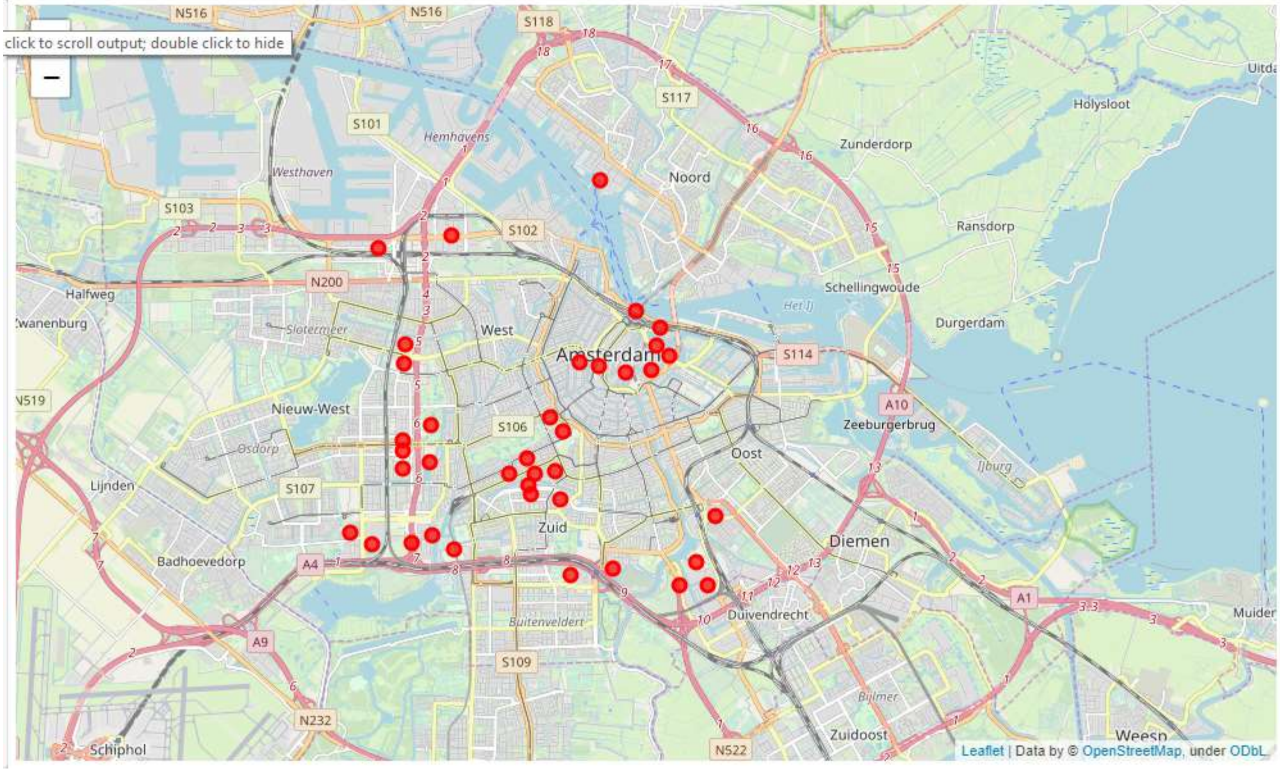
```
show_clusters(istanbul_merged, lat_istanbul, long_istanbul, 12)
```



Visualization of similar neighbourhoods in Amsterdam;

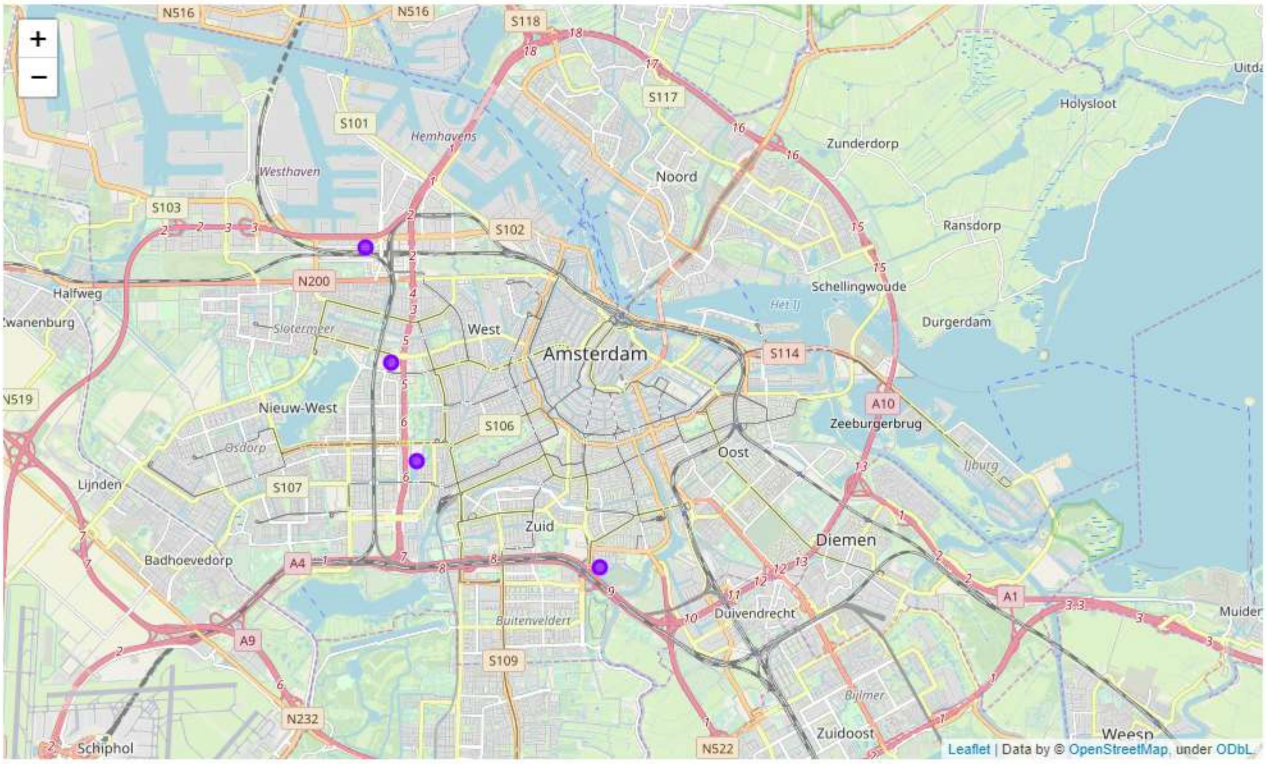


```
show_clusters(cluster_0, lat_amsterdam, long_amsterdam, 12)
```



Neighbourhoods that have Turkish venues;

```
show_analised(cluster_0_analised, lat_amsterdam, long_amsterdam, 12)
```



6. Conclusion

Purpose of this project was to identify neighbourhoods in similar with your home town and keep your adaptation period as short as possible. For a newcomer the hardest thing to find similar taste of your culture in vicinity.By getting venues information from Foursquare data we have first clustered neighbourhoods, and then by filtering the collection of locations which is similar to home town regarding existing turkish venues in vicinity we reached 4 neighbourhoods locations. We presented them on the map.

You can decide optimal neighbourhood location to move on based on our analysis.