Battle of Neighborhoods - Gyms in Amsterdam

Clustering City Districts for Potential Gym Locations

By Tiina Vaahtio

08.01.2020, Amsterdam, Netherlands

1. Introduction

In the recent years the fitness industry has been booming in the Netherlands. New sports locations, apps, blogs, instagram influencers related to fitness and well-being keep appearing at a fast pace and many gym chains are growing and expanding with a high rate. According to CBS, in 2017, 2.25 million people in the Netherlands had a gym membership. Moreover, on avg a fitness centers have over 1100 members.

In the Netherlands Gym and fitness industry is mostly controlled by gym-chains, with Basic-Fit being the biggest with 152 locations (2018), and SportsCity/FitForFree falling second with 107 locations (Statista, 2018)

With a high trending industry competition is also fierce. Moreover, property cost in Amsterdam are scoring at all time high. Hence, opening up a new location comes with high fixed cost and a big risk. Therefore, before opening up a new location, a gym chain will have to carry out extensive analysis on the best neighborhood to open up a new facility. This project will provide one way to analyze different neighborhoods based on the number of gyms, their ratings and population sizes. The analysis can support decision making on which neighborhoods have a high potential for a new gym location. Moreover, we can identify neighborhoods that are already saturated and don't have a good potential for expansion.

2. Data

To successfully build this project and understand the best neighborhoods to open up a gym in Amsterdam we will be using the following data sources:

- Amsterdam Neighborhood Data

- Data Source: https://claircitydata.cbs.nl/dataset/districts-and-neighbourhoods-amsterdam/resource/d02c5f12-1cfa-4d7c-91d3-41af8e4ed634?view_id=5bae9d1c-2bfc-4f00-b15a-69a27632c41c
- Description: This dataset describes the different neighborhoods in Amsterdam with their respective latitudes and longitudes. It also contains useful information about the population of different age groups in each neighborhood. This dataset also contains the GEOJSON features that will be used to visualize the data on a choropleth map.

- Gyms/Fitness Centers in Amsterdam Neighborhoods

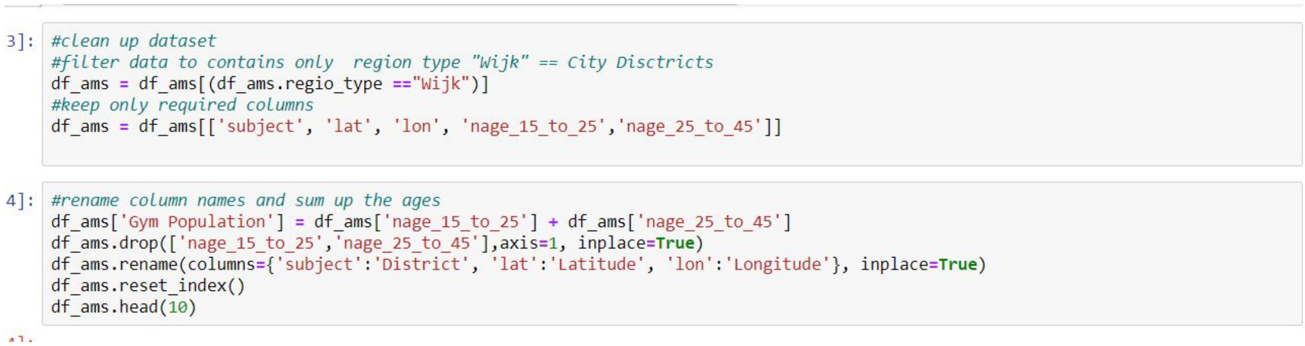
- Data Source: FourSquare API
- Description: FourSquare API can provide us venue information that we can retrieve based on the Amsterdam Neighborhood latitudes and longitudes. We will further limit this data to include only information about sports venues (gyms) in Amsterdam Area. We will retrieve information such as: Venue Name, Category, Location, Rating, Likes etc.

2. 1 Data Preparation

As in any typical data analytics project, most of the time is spent on data preparation. Initially, I was on the lookout on Amsterdam neighborhood data with city district. The city of Amsterdam has great open data sources to look for, however the challenge to find consistent data with both geo-information as well as population with the limited local language knowledge, was a tedious job. However, eventually I managed to find exactly what I was looking for.

The first step was to clean-up the dataset. This included steps like dropping unnecessary rows, creating derived columns, and formatting the geojson information to a geojson file. The cleaning and the transformations were carried out in my jupyter notebook. However, the geojson file I manually reformatted in Atom.App-code editor.

The city district data set included information also from the population. After some background study, I decided to calculate the populations from 15 - 45 years old, as these are the most likely age groups to have gym membership.



```
3]: #clean up dataset
    #filter data to contains only  region type "Wijk" == City Disctricts
    df_ams = df_ams[(df_ams.regio_type =="Wijk")]
    #keep only required columns
    df_ams = df_ams[['subject', 'lat', 'lon', 'nage_15_to_25','nage_25_to_45']]
```

```
4]: #rename column names and sum up the ages
    df_ams['Gym Population'] = df_ams['nage_15_to_25'] + df_ams['nage_25_to_45']
    df_ams.drop(['nage_15_to_25','nage_25_to_45'],axis=1, inplace=True)
    df_ams.rename(columns={'subject':'District', 'lat':'Latitude', 'lon':'Longitude'}, inplace=True)
    df_ams.reset_index()
    df_ams.head(10)
```

Image 1: Data Preparation example

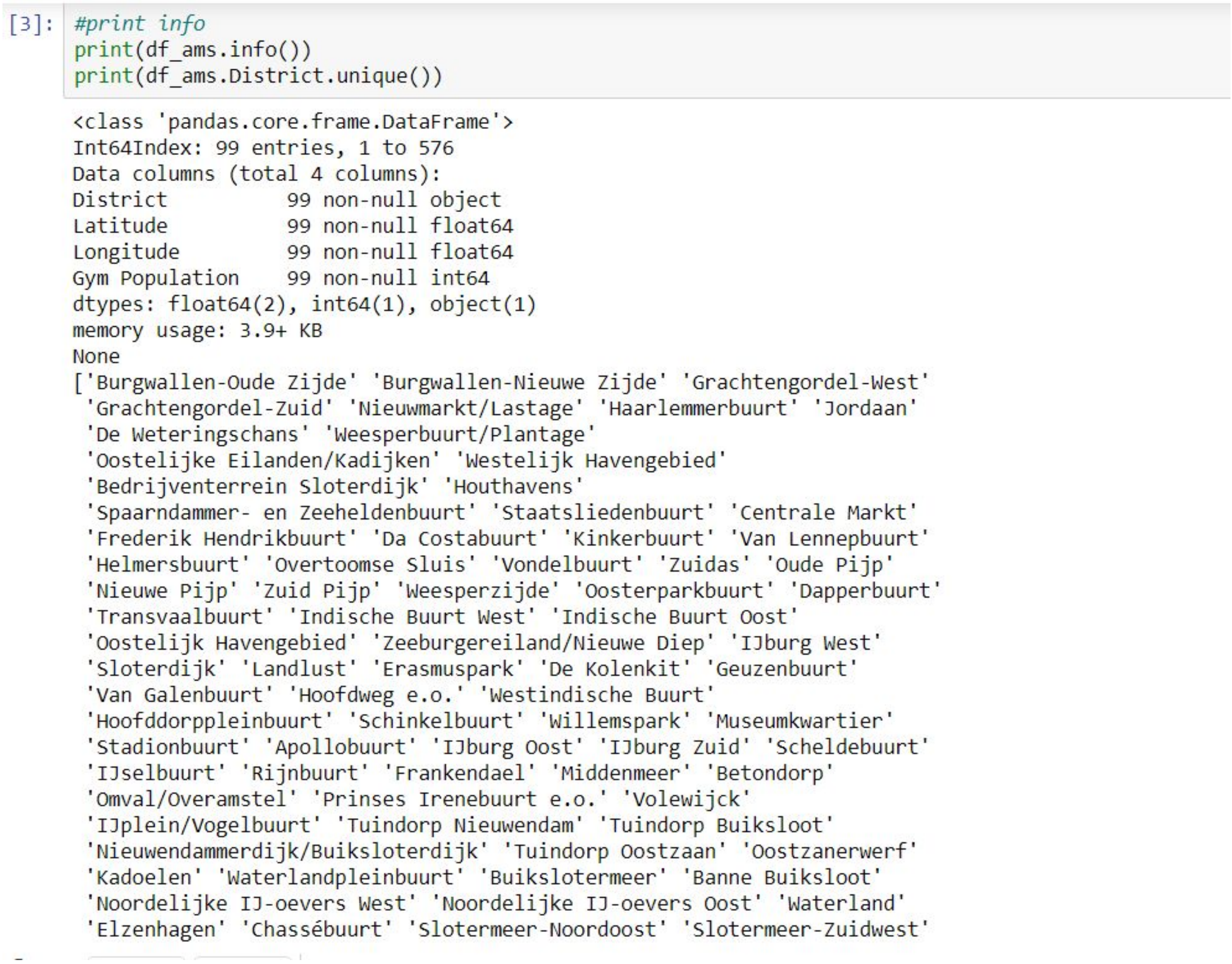
The second part was to retrieve information from the Foursquare API. Foursquare is one of the largest location-based data providers, and it was part of the Capstone course assignment to learn how retrieve data via APIs.

In this project, the Foursquare API, was used in two ways. First, to get venue locations and categories, per the districts in Amsterdam. Secondly, this venue data was used to retrieve venue details such as Rating and Likes.

Eventually, all these three datasets were merged into one dataset grouped by districts, and filtered with venu category that related to Gym / Fitness, to carry out the analysis.

3. Data Exploration

The analysis were started by exploring some of the basics of the datasets. For example, types, unique values, shapes etc.



```
[3]: #print info
     print(df_ams.info())
     print(df_ams.District.unique())

     <class 'pandas.core.frame.DataFrame'>
     Int64Index: 99 entries, 1 to 576
     Data columns (total 4 columns):
     District          99 non-null object
     Latitude          99 non-null float64
     Longitude         99 non-null float64
     Gym Population    99 non-null int64
     dtypes: float64(2), int64(1), object(1)
     memory usage: 3.9+ KB
     None
     ['Burgwallen-Oude Zijde' 'Burgwallen-Nieuwe Zijde' 'Grachtengordel-West'
      'Grachtengordel-Zuid' 'Nieuwmarkt/Lastage' 'Haarlemmerbuurt' 'Jordaan'
      'De Weteringschans' 'Weesperbuurt/Plantage'
      'Oostelijke Eilanden/Kadijken' 'Westelijk Havengebied'
      'Bedrijventerrein Sloterdijk' 'Houthavens'
      'Spaarndammer- en Zeeheldenbuurt' 'Staatsliedenbuurt' 'Centrale Markt'
      'Frederik Hendrikbuurt' 'Da Costabuurt' 'Kinkerbuurt' 'Van Lennepbuurt'
      'Helmersbuurt' 'Overtoomse Sluis' 'Vondelbuurt' 'Zuidas' 'Oude Pijp'
      'Nieuwe Pijp' 'Zuid Pijp' 'Weesperzijde' 'Oosterparkbuurt' 'Dapperbuurt'
      'Transvaalbuurt' 'Indische Buurt West' 'Indische Buurt Oost'
      'Oostelijk Havengebied' 'Zeeburgereiland/Nieuwe Diep' 'IJburg West'
      'Sloterdijk' 'Landlust' 'Erasmuspark' 'De Kolenkit' 'Geuzenbuurt'
      'Van Galenbuurt' 'Hoofdweg e.o.' 'Westindische Buurt'
      'Hoofddorppleinbuurt' 'Schinkelbuurt' 'Willemspark' 'Museumkwartier'
      'Stadionbuurt' 'Apollobuurt' 'IJburg Oost' 'IJburg Zuid' 'Scheldebuurt'
      'IJselbuurt' 'Rijnbuurt' 'Frankendael' 'Middenmeer' 'Betondorp'
      'Omval/Overamstel' 'Prinses Irenebuurt e.o.' 'Volewijck'
      'IJplein/Vogelbuurt' 'Tuindorp Nieuwendam' 'Tuindorp Buiksloot'
      'Nieuwendammerdijk/Buiksloterdijk' 'Tuindorp Oostzaan' 'Oostzanerwerf'
      'Kadoelen' 'Waterlandpleinbuurt' 'Buikslotermeer' 'Banne Buiksloot'
      'Noordelijke IJ-oevers West' 'Noordelijke IJ-oevers Oost' 'Waterland'
      'Elzenhagen' 'Chassébuurt' 'Slotermeer-Noordoost' 'Slotermeer-Zuidwest'
```

Image 2: Data Exploring Examples

From the foursquare data the following details were initially gathered:

- 99 unique districts
- 326 unique venue categories

After limiting the venue dataset to Gym / Fitness related categories eventually, we had a dataset on venue level with 68 venue details.

These gym locations were visualized on a map:

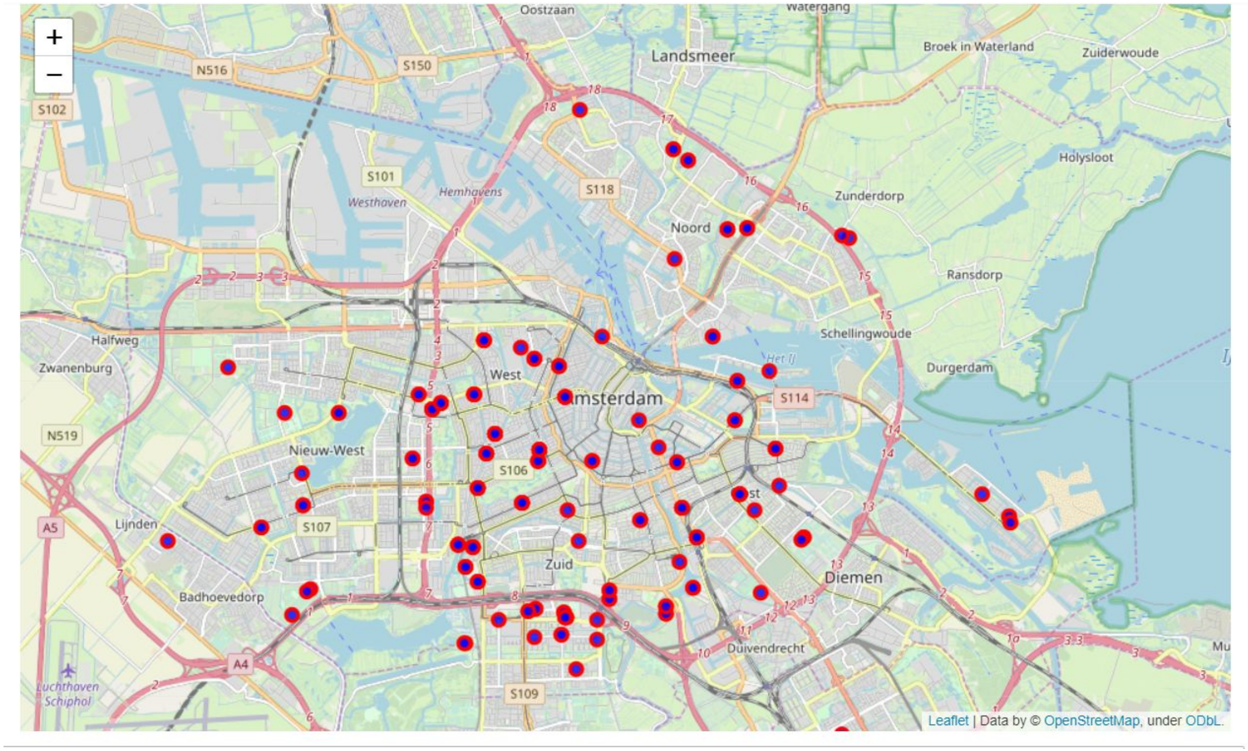
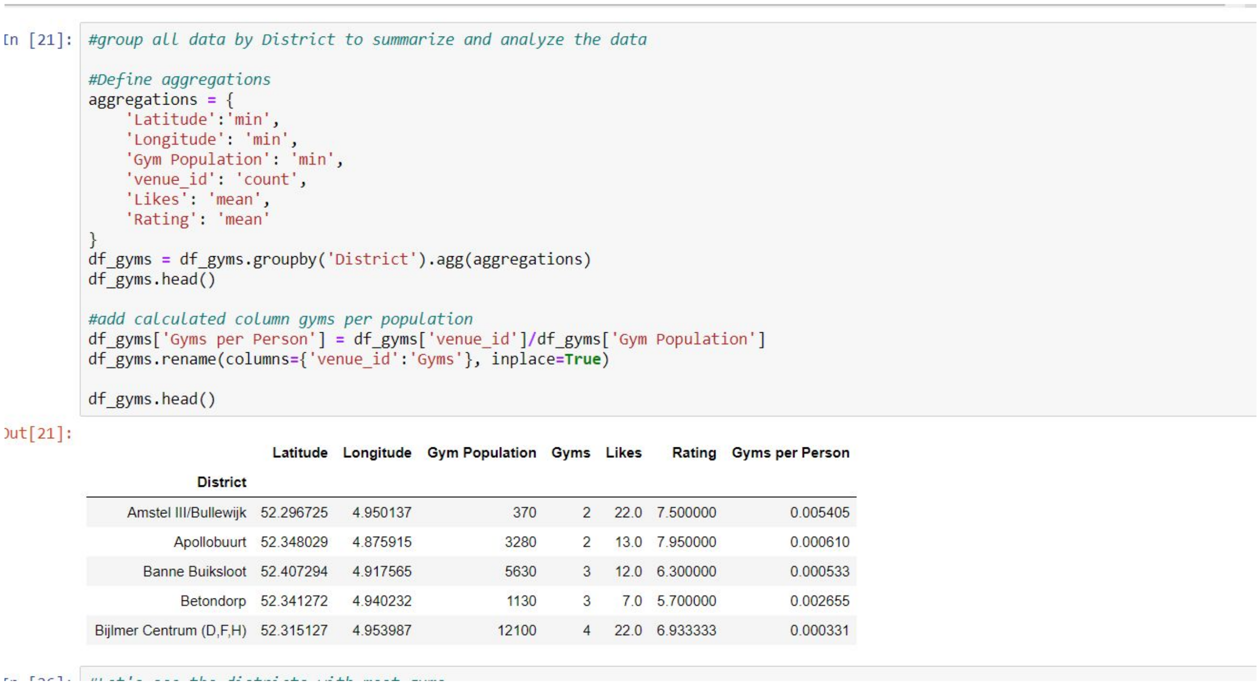


Image 3: Gym Locations in Amsterdam

After these individual location analysis, the dataset was grouped by District and some aggregated values were retrieved.



```
In [21]: #group all data by District to summarize and analyze the data

         #Define aggregations
         aggregations = {
             'Latitude':'min',
             'Longitude': 'min',
             'Gym Population': 'min',
             'venue_id': 'count',
             'Likes': 'mean',
             'Rating': 'mean'
         }
         df_gyms = df_gyms.groupby('District').agg(aggregations)
         df_gyms.head()

         #add calculated column gyms per population
         df_gyms['Gyms per Person'] = df_gyms['venue_id']/df_gyms['Gym Population']
         df_gyms.rename(columns={'venue_id':'Gyms'}, inplace=True)

         df_gyms.head()
```

Out[21]:

District	Latitude	Longitude	Gym Population	Gyms	Likes	Rating	Gyms per Person
Amstel III/Bullewijk	52.296725	4.950137	370	2	22.0	7.500000	0.005405
Apollobuurt	52.348029	4.875915	3280	2	13.0	7.950000	0.000610
Banne Buiksloot	52.407294	4.917565	5630	3	12.0	6.300000	0.000533
Betondorp	52.341272	4.940232	1130	3	7.0	5.700000	0.002655
Bijlmer Centrum (D,F,H)	52.315127	4.953987	12100	4	22.0	6.933333	0.000331

Image 4: Aggregated dataset

From the grouped dataset, the following exploratory analysis were carried out:

- Top 15 Districts
  - By number of gyms
  - By average rating
  - By gyms per population

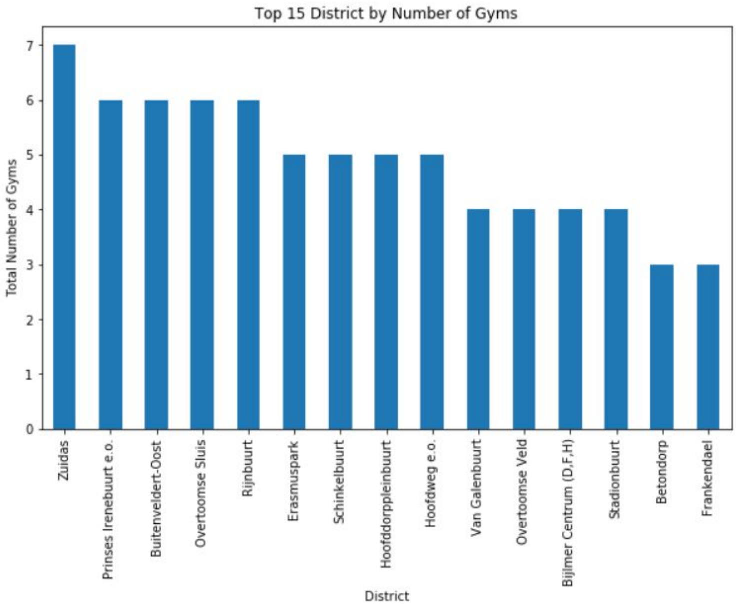


Image 5: Top Districts by number of gyms

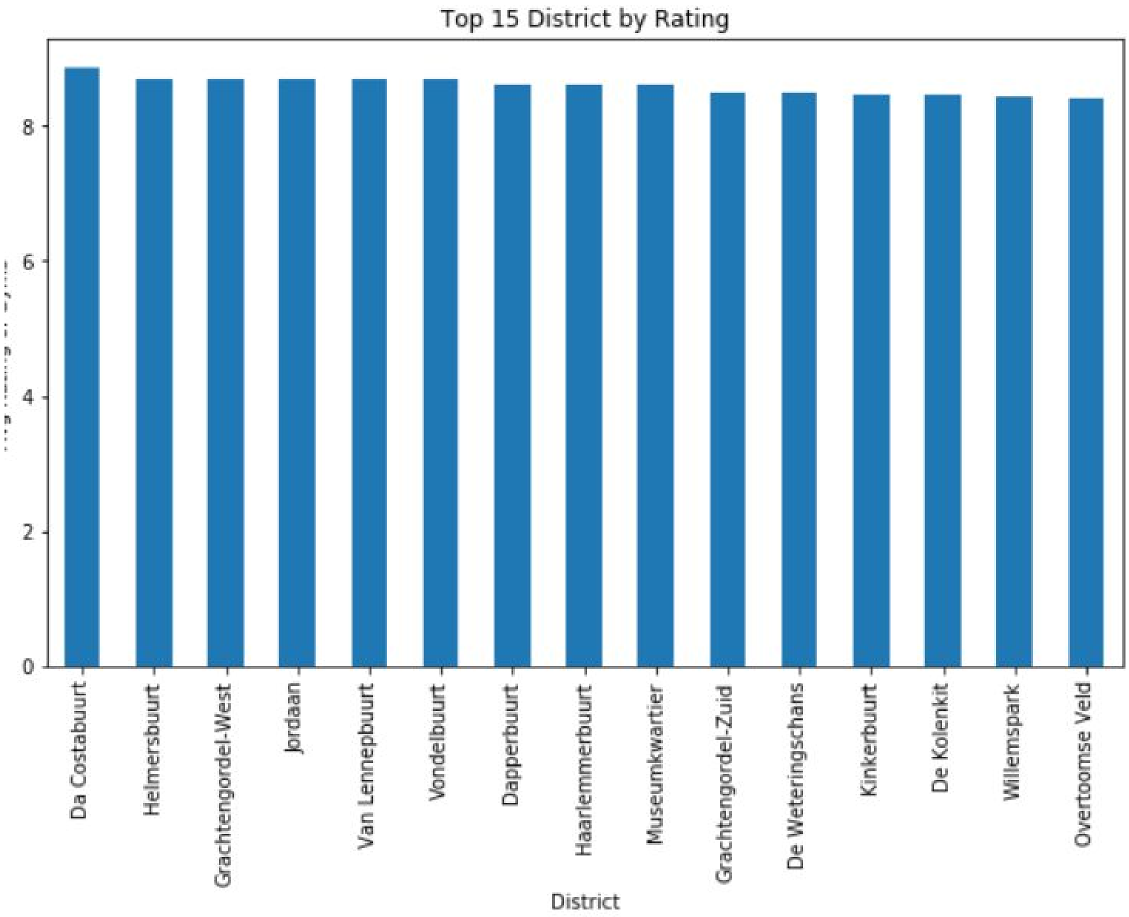


Image 6: Districts by highest rating

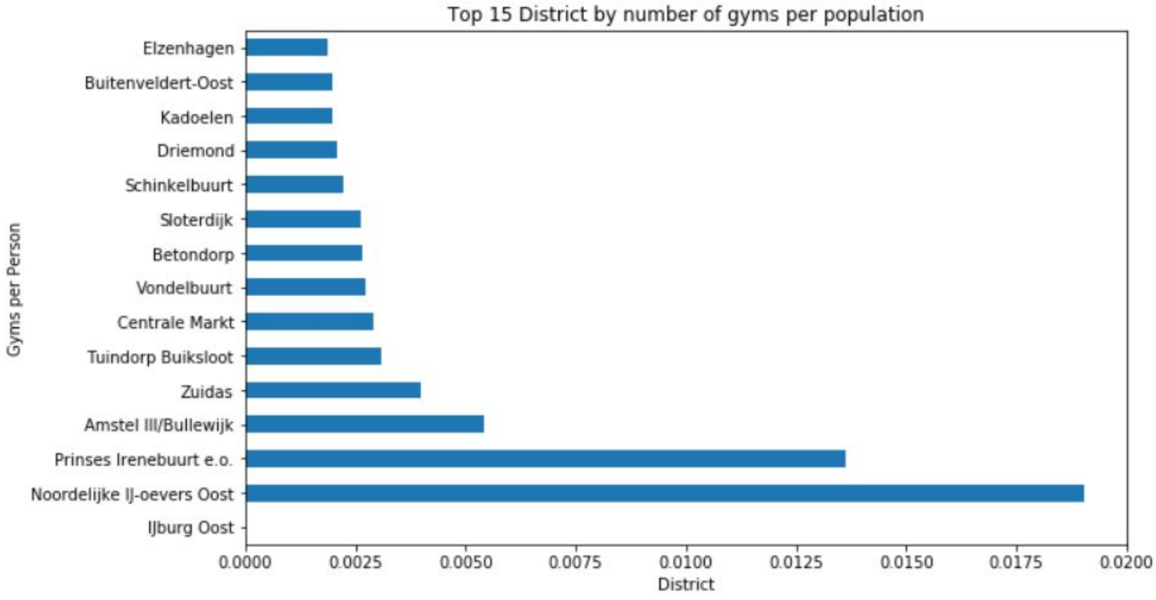


Image 7: Districts by highest number of gyms per population
This data was also explored on the map visualization.

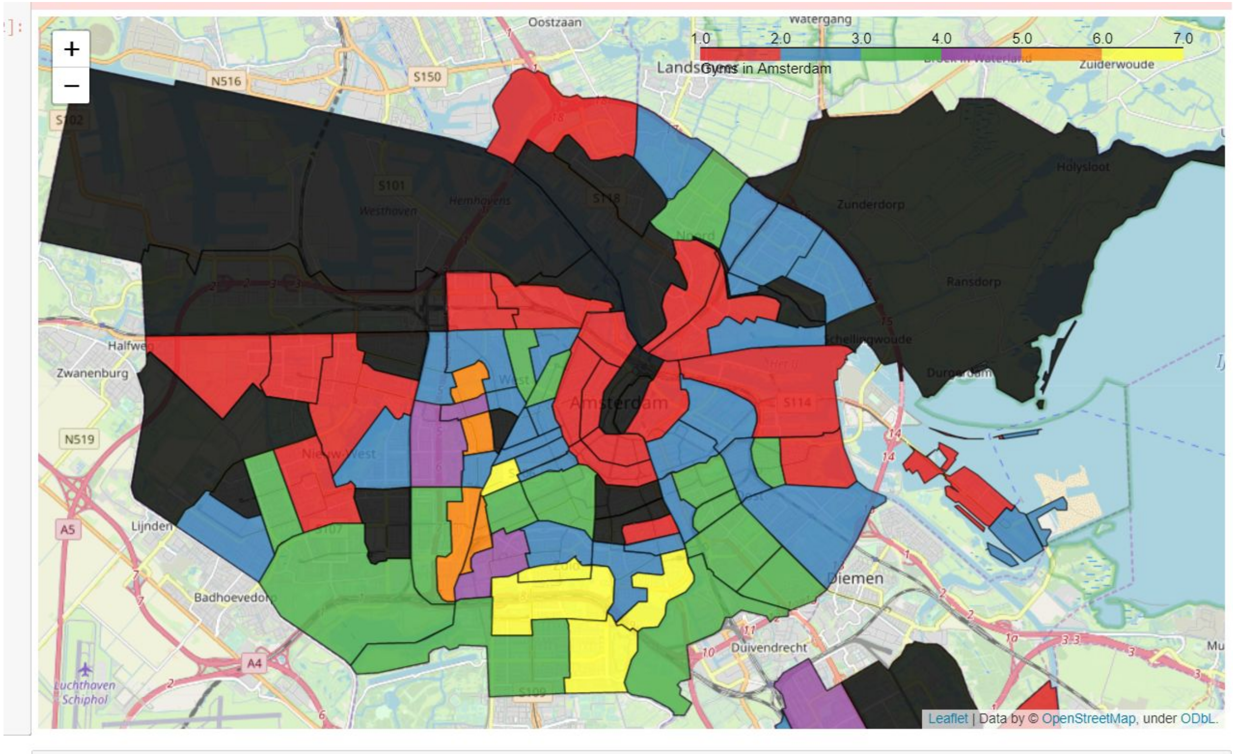


Image 8: Number of gyms per district

4. Clustering

The final part of the analysis was to use k-means clustering to define the neighborhoods to similar groups.

The clustering process was carried out with iteration, exploring different number of clusters. The values were also normalized using the standard scaler to carry out the analysis. Eventually, including 5 clusters, was descriptive enough to categorize the districts.

The explorative analysis of the clusters were carried out after the clustering modeling. Initially, first the different clusters were visualized on a map.

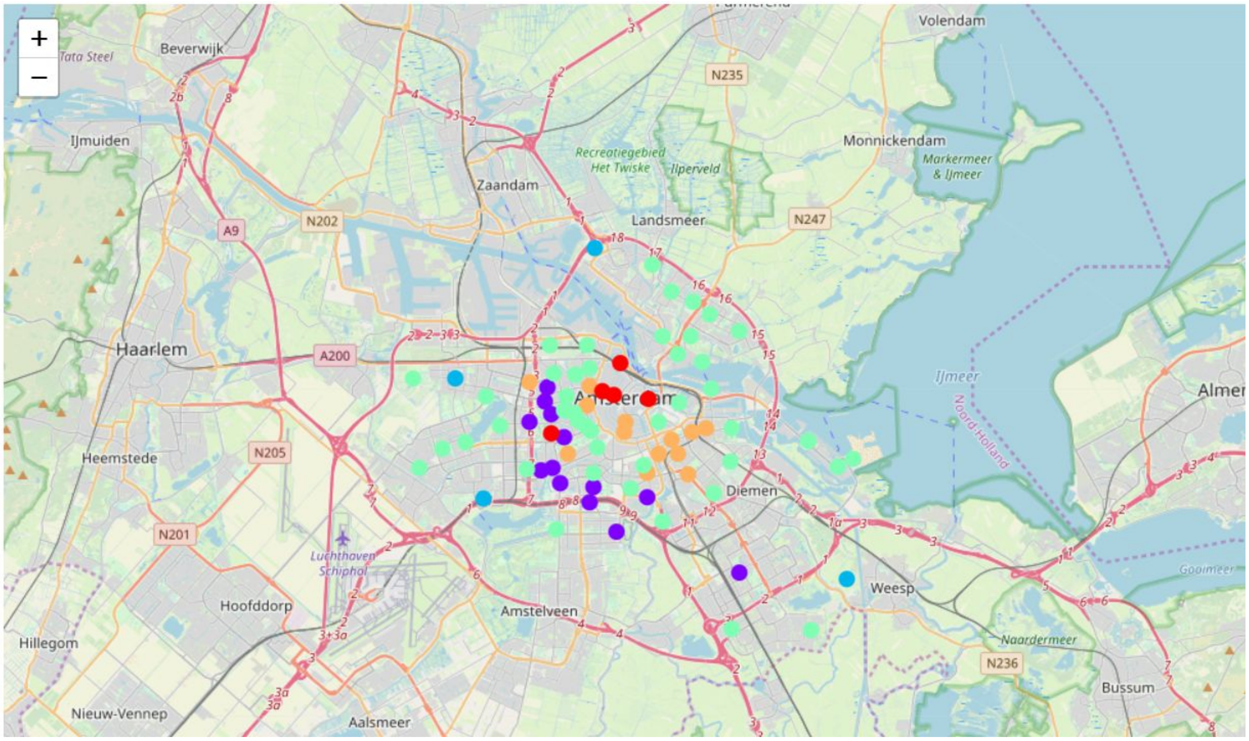


Image 9: Clusters of Amsterdam districts

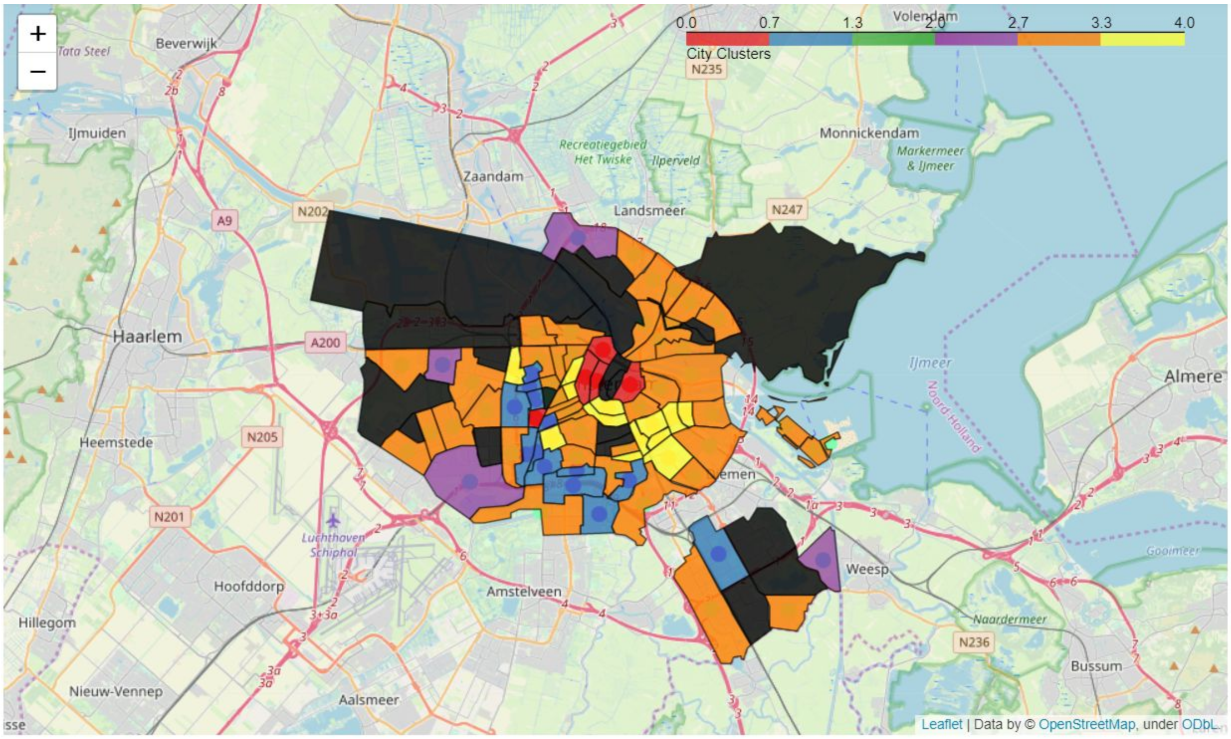


Image 10: Clusters on Choropleth

After the overall analysis of the clustering, each of the clusters were taken into individual analysis, in order to characterize and describe them meaningfully. Moreover, these individual analysis were the key in order to draw conclusions and recommendations about the data and analysis.

Eventually the districts were categorized to 5 categories as follows:
(High Pop= High Population, Gyms pp = Gyms per person in the district)

- High Pop/High Rating/Low Gyms pp
- High Pop/Med Rating/High Gyms pp
- Med Pop/No Ratings/Med Gyms pp
- Med Pop/Low Rating/Med Gyms pp
- Med Pop/High Rating/Low Gym pp

5. Discussion and Conclusions

These analysis carried out on the gym location and customer rating information can help out to determine potential new locations for gym chains to determine where to open up a new facility. It can help to analyze which districts are already populated, where are the most satisfied customers, and which neighborhoods are similar.

Based on the analysis and the k-means clustering of the city districts of Amsterdam, 5 type of descriptive characteristics for were identified. By understanding, the cluster the neighborhood belongs to, the gym chains can make quick initial analysis, whether they should look into more detail to specific neighborhoods for new potential locations.

I recommend that especially two of the cluster types could be considered as a potential starting place. These two clusters are:

- High Pop/High Rating/Low Gyms pp
- Med Pop/High Rating/Low Gym pp

Both of these clusters have still room for new gym locations, on average the gym-goers are satisfied with the gym services and moreover, there is a high potential gym population to gain customers from.

Secondly, I would advise to look into the “Med Pop/Low Rating/Med Gyms pp”, as these districts are not yet fully saturated and their customers have a low rating on the existing gyms. A new player can open up in these districts and dominate the game with a superior experience.

Of course as with any analysis, further analysis and additional data could be used to make more conclusive decisions. Studies have shown that gym-membership is heavily linked on education and income levels, and retrieving this information and combining with the current analysis could be carried out to gain even an better understanding of the suitability of the districts.

6. Limitations

- These analysis are limited by the nature of data used in the analysis
- The accuracy of the gym details is depended on the data available from the Foursquare API
- The carrier of this project is a beginner in Python and the analysis are still limited to superficial level

8. Sources

"Trends in the Netherlands", Figures Leisure,
https://longreads.cbs.nl/trends18-eng/society/figures/leisure/

"Leading Fitness Centers in the Netherlands, by number of clubs (2018)
https://www.statista.com/statistics/819238/leading-fitness-centers-in-the-netherlands-by-number-of-clubs/