OPENING A SECOND GYM LOCATION

Wait but *where?*

TABLE OF CONTENTS

1. INTRODUCTION

The owners of a gym are looking to open their second location. Their current gym is located in the borough Feijenoord, Rotterdam, The Netherlands.
Business is going well. Their concept is mainly focussed on the youth and through initiatives with local restaurants, bars, and sport clubs the gym is advertised. This is thought to be a major part of the success, and makes the location of the gym crucial.
The second location is to be opened in Amsterdam, the Netherlands. The owners are not familiar with this area, because of the long lasting rivalry between the two cities they are yet to visit the place! Not wanting to spend more time in Amsterdam then absolutely necessarily, they decide to rely on data to find the boroughs most similar to Feijenoord.

1.1 The business problem

The concept of the current gym is heavily depended on the location. To ensure success of the second location in Amsterdam, the area should be similar to the current location. However, it is unknown which boroughs in Amsterdam are similar to Feijenoord. Therefore, the goal of the project is to find the answer to the following research question: "Which of the boroughs in Amsterdam are similar to Feijenoord, Rotterdam based on the target group of the gym (youths) and the presence of local enterprises (e.g. restaurants, bars and sport clubs)?

1.2. The Analytical Approach

To answer the research question the neighbourhoods should be grouped. A clustering approach is best suited as the data is unlabelled. The boroughs in the same group as Feijenoord are similar to the current area of the gym. An example of a clustering algorithm which can be used is k-means.

2. DATA REQUIREMENTS AND SOURCES

First the different boroughs in Amsterdam should be mapped. In order for data to be relevant it must be possible to group the data according to these boroughs. Furthermore, the data should be recently updated. For a relevant comparison it is crucial to only include relevant data. This is based on the research question and include:

- Population composition based on age and income
- Number of local restaurants, bars and sport clubs relative to the total population

This means the data is two-fold. The first part consist out of local population composition data. The second part is purely location data.

2.1 Population composition

In The Netherlands the Central Bureau of Statistics (CBS) is the national statical office which provided reliable statical information and data. (Much of) the data is open source and can be downloaded in csv format from their data portal. This is includes a dataset Key figures of boroughs and neighbourhoods:

https://opendata.cbs.nl/statline/#/CBS/nl/dataset/84799NED/table?ts=1612255554791

which provides key figures such as population compositions per neighbourhood and borough overall for all major cities in The Netherlands. Metadata of the dataset can be found in this csv file:

https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS&tableId=84799NED&_theme=235#

Besides providing descriptions of the different columns in the dataset, the metadata also links the "Codering_3" column to the names of the boroughs. Figure 3 shows a snapshot of a part of the dataset and of the metadata.

ID	WijkenEnBuurten	Gemeentenaam_1	SoortRegio_2	Codering_3	IndelingswijzigingWijkenEnBuurten_4	AantalInwoners_5	Mannen_6	Vrouwen_7	k_0Tot15Jaar_8	k_15Tot25Jaar_9	k_25Tot45Jaar_10	k_45Tot65Jaar_11	k_65JaarOfOuder_12
921	GM0363	Amsterdam	Gemeente	GM0363	.	872757	432879	439878	125119	108636	317667	210376	110959
922	WK036300	Amsterdam	Wijk	WK036300	1	4465	2465	2000	220	695	2195	940	425
923	BU03630000	Amsterdam	Buurt	BU03630000	1	1100	625	470	50	125	575	280	80
924	BU03630001	Amsterdam	Buurt	BU03630001	1	730	410	315	30	135	370	135	70
925	BU03630002	Amsterdam	Buurt	BU03630002	1	1610	855	755	90	260	780	315	180
926	BU03630003	Amsterdam	Buurt	BU03630003	1	350	195	155	20	60	180	70	30
927	BU03630004	Amsterdam	Buurt	BU03630004	1	670	370	300	35	125	290	150	80
928	WK036301	Amsterdam	Wijk	WK036301	1	4130	2255	1875	195	615	2220	785	335
929	BU03630100	Amsterdam	Buurt	BU03630100	1	0	0	0	0	0	5	0	0
930	BU03630101	Amsterdam	Buurt	BU03630101	1	575	320	255	25	85	320	115	45
931	BU03630102	Amsterdam	Buurt	BU03630102	1	460	270	190	20	85	255	95	25
932	BU03630103	Amsterdam	Buurt	BU03630103	1	825	445	375	35	115	450	165	70
933	BU03630104	Amsterdam	Buurt	BU03630104	1	820	460	355	55	130	460	140	45
934	BU03630105	Amsterdam	Buurt	BU03630105	1	720	400	315	45	90	375	155	70
935	BU03630106	Amsterdam	Buurt	BU03630106	1	415	165	245	15	55	210	80	75
936	BU03630107	Amsterdam	Buurt	BU03630107	1	305	180	125	15	70	165	45	20
937	WK036302	Amsterdam	Wijk	WK036302	1	6435	3370	3065	470	870	2265	1745	1100
938	BU03630200	Amsterdam	Buurt	BU03630200	1	1870	975	890	175	200	600	510	395
939	BU03630201	Amsterdam	Buurt	BU03630201	1	1820	925	895	125	360	630	445	270
940	BU03630202	Amsterdam	Buurt	BU03630202	1	2055	1105	950	130	225	790	605	320
941	BU03630203	Amsterdam	Buurt	BU03630203	1	685	360	325	45	85	255	185	120
942	WK036303	Amsterdam	Wijk	WK036303	1	5430	2925	2505	460	720	2090	1370	800
943	BU03630300	Amsterdam	Buurt	BU03630300	1	820	440	375	85	80	360	195	110
944	BU03630301	Amsterdam	Buurt	BU03630301	1	145	80	65	15	25	55	35	25
945	BU03630302	Amsterdam	Buurt	BU03630302	1	1415	725	685	120	195	460	390	255
946	BU03630303	Amsterdam	Buurt	BU03630303	1	1510	790	720	150	205	465	450	250
947	BU03630304	Amsterdam	Buurt	BU03630304	1	485	285	200	25	55	290	85	40
948	BU03630305	Amsterdam	Buurt	BU03630305	1	385	215	170	20	90	225	50	10

NL00	Nederland	
GM1680	Aa en Hunze	Gemeente 'Aa en Hunze' (GM1680)
WK168000	Wijk 00 Annen	Wijk 'Wijk 00 Annen' (WK168000) is een wijk in de gemeente 'Aa en Hunze'.
BU16800000	Annen	Buurt 'Annen' (BU16800000) is een buurt in wijk 'Wijk 00 Annen' in de gemeente 'Aa en Hunze'.
BU16800009	Verspreide huizen Annen	Buurt 'Verspreide huizen Annen' (BU16800009) is een buurt in wijk 'Wijk 00 Annen' in de gemeente 'Aa en Hunze'.
WK168001	Wijk 01 Eext	Wijk 'Wijk 01 Eext' (WK168001) is een wijk in de gemeente 'Aa en Hunze'.
BU16800100	Eext	Buurt 'Eext' (BU16800100) is een buurt in wijk 'Wijk 01 Eext' in de gemeente 'Aa en Hunze'.
BU16800109	Verspreide huizen Eext	Buurt 'Verspreide huizen Eext' (BU16800109) is een buurt in wijk 'Wijk 01 Eext' in de gemeente 'Aa en Hunze'.
WK168002	Wijk 02 Anloo	Wijk 'Wijk 02 Anloo' (WK168002) is een wijk in de gemeente 'Aa en Hunze'.
BU16800200	Anloo	Buurt 'Anloo' (BU16800200) is een buurt in wijk 'Wijk 02 Anloo' in de gemeente 'Aa en Hunze'.
BU16800209	Verspreide huizen Anloo	Buurt 'Verspreide huizen Anloo' (BU16800209) is een buurt in wijk 'Wijk 02 Anloo' in de gemeente 'Aa en Hunze'.
WK168003	Wijk 03 Gasteren	Wijk 'Wijk 03 Gasteren' (WK168003) is een wijk in de gemeente 'Aa en Hunze'.
BU16800300	Gasteren	Buurt 'Gasteren' (BU16800300) is een buurt in wijk 'Wijk 03 Gasteren' in de gemeente 'Aa en Hunze'.
BU16800309	Verspreide huizen Gasteren	Buurt 'Verspreide huizen Gasteren' (BU16800309) is een buurt in wijk 'Wijk 03 Gasteren' in de gemeente 'Aa en Hunze'.
WK168004	Wijk 04 Anderen	Wijk 'Wijk 04 Anderen' (WK168004) is een wijk in de gemeente 'Aa en Hunze'.
BU16800400	Anderen	Buurt 'Anderen' (BU16800400) is een buurt in wijk 'Wijk 04 Anderen' in de gemeente 'Aa en Hunze'.
BU16800409	Verspreide huizen Anderen	Buurt 'Verspreide huizen Anderen' (BU16800409) is een buurt in wijk 'Wijk 04 Anderen' in de gemeente 'Aa en Hunze'.
WK168005	Wijk 05 Schipborg	Wijk 'Wijk 05 Schipborg' (WK168005) is een wijk in de gemeente 'Aa en Hunze'.
BU16800500	Schipborg	Buurt 'Schipborg' (BU16800500) is een buurt in wijk 'Wijk 05 Schipborg' in de gemeente 'Aa en Hunze'.

Figure 1: Snapshot of the CBS data set and the corresponding meta data

As the dataset includes about 100 columns, table 1 below only describes the relevant columns based on the research question. Almost all data will be from 2020. The only exception is the average income, the most recent dataset for which this column is complete is 2017.

In 5 columns the population per age interval is given, together with the column total population this can be used to create a age distribution for the different boroughs. GemiddeldInkomenPerInwoner_66 gives the average income per borough. This data in not available yet for 2020. The most recent data is in the data set for 2017 and will be used instead.

Table 1: Description of relevant columns in CBS dataset

Column names	Descriptions
Gemeentenaam_1	Gives the name of the corresponding city of the neighbourhood or borough
WijkenEnBuurten	Gives the code of the neighbourhoods and boroughs, the metadata file can be used to retrieve the name using this key
SoortRegio_2	Gives the type of area: Gemeente: township, Wijk: borough, Buurt: neighbourhood
AantalInwoners_5	Gives the total population of the neighbourhoods or borough
k_0Tot15Jaar_8 ... k_65JaarOfOuder	Population between intervals of age: 0 - 15 years, 15 to 25 years, 25 to 45 years, 45 to 65 years, 65 or older
GemiddeldInkomenPerInwoner_66	Average income per Person in thousands, not available for 2020, 2017 used instead

2.2 Location data

Information about the type and number of local enterprises, such as restaurants, bars, and sport clubs will be acquired using the location data of Foursquare. In order to retrieve information from Foursquare the coordinates of the boroughs are needed. These coordinates will be retrieved using the package geopy, example given below. The name of the neighbourhoods will be retrieved from the "Key figures of boroughs and neighbourhoods" metadata file. The number of enterprises per type will be retrieved from Foursquare and normalized to the total population in that area. This can then be used to compare the neighbourhoods together with the population compositions.

2.3 Data cleaning

Reading in the CBS csv files

The 2020 CBS dataset was imported using the pandas.read_csv function. First, the relevant columns given in table 1, excluding the average income, were extracted. The 2017 dataset was imported in a similar way, only extracting the "WijkenEnBuurten" and "GemiddeldInkomenPerInwoner_66" columns. These two data sets were merged on the "WijkenEnBuurten" key.
The dataset included total city and neighbourhood data while only the borough data was deemed relevant. Therefore, the data was filtered on the 'SoortRegio_2" column only including the rows with the "Wijk" as value. The columns were renamed to English names:

- WijkenEnBuurten: Key
- Gemeentenaam_1: City
- SoortRegio_2: Area_type
- AantalInwoners_5: Total Population
- k_0Tot15Jaar_8: 0_to_15_years
- k_15Tot25Jaar_9: 15_to_25_years
- k_25Tot45Jaar_10: 25_to_45_years
- k_45Tot65Jaar_11: 45_to_65_years
- k_65JaarOfOuder_12: 65_years_or_older
- GemiddeldInkomenPerInwoner_66: Average_income_2017

The population per age interval was given in absolute numbers, making comparisons down the road difficult. Therefore, these are normalized with the total population in the neighbourhood, giving the age distribution in intervals.

Next the names of the neighbourhoods were added to data frame. These were given in the metadata csv file, together with the Key. By reading in the csv file and merging on the key, the names were added. This lead to data frame with 98 boroughs in Amsterdam and 17 in Rotterdam. All boroughs in Rotterdam but Feijenoord were dropped.

Adding location data
Before getting the location data for the different neighbourhoods from Foursquare, the latitudes and longitudes for the neighbourhoods are needed. No location data could be found for the neighbourhoods. Therefore, Geopy package was used. This package retrieves the latitudes and longitudes based on the address. First a address column was created using the name of the neighbourhood and the city. Not all addresses are found, for these address a 0.00 was returned. This was the case for 25 neighbourhoods. No other database was found to get the coordinates for these neighbourhoods, so they were dropped.

The venues in a radius of 1000m around the latitudes and longitudes were retrieved from Foursquare. This resulted in a dataframe in which every row represented a venue and the column gave the longitude,

latitude, name, and category. Figure 2 shows a snapshot of the dataframe. The categories relevant for the research questions were: gyms, restaurants, bars, and sport schools. However, the categories were more varied, i.e. also giving the type of restaurants. Therefore, the categories containing the word "restaurant were renamed to 'RESTAURANT'. Similar processing was performed for the other relevant categories.

[40]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
2056	Feijenoord	51.909661	4.506957	Restaurant Allure	51.909769	4.499476	RESTAURANT
2061	Feijenoord	51.909661	4.506957	Safir	51.910683	4.499503	RESTAURANT

Figure 2: Snapshot of the Venues dataframe

One-hot encoding was used for Venue Category and the data was grouped by the neighbourhood and counted. Finally, the CBS data and the foursquare data was merged and the number of venues was normalized to the population in a borough.

3. METHODOLGY

3.1 Exploratory Data Analysis

Before starting to model the data the types, distributions, min and max values of the different variables were investigated. All the data types are as expected, all the relevant variables for modelling are either floats or integers.
Table 2 shows the descriptive statistics of the variables. In all cases the mean, min, max, and std of the variables seem to be logical. There does seem to be large differences between the means and medians of the variable. Therefore, normalization of the data set is necessary before modelling.

Table 2: Descriptive statistics of data table

Variable	mean	std	min	25%	50%	75%	max
Total Population	9700	9277	535	5133	8270	12075	76590
0_to_15_years	0.14	0.05	0.02	0.11	0.13	0.17	0.29
15_to_25_years	0.12	0.05	0.06	0.10	0.11	0.13	0.50
25_to_45_years	0.38	0.09	0.18	0.30	0.37	0.44	0.64
45_to_65_years	0.24	0.05	0.09	0.22	0.24	0.27	0.34
65_years_or_older	0.13	0.05	0.02	0.09	0.13	0.16	0.24
Average_income_2017	31	11	18	23	29	36	69
NumOfBars	4.86E-04	5.45E-04	0.00E+00	1.80E-04	3.34E-04	5.50E-04	2.70E-03
NumOfGyms	5.90E-05	1.07E-04	0.00E+00	0.00E+00	0.00E+00	8.40E-05	5.65E-04
NumOfRestaurants	1.27E-03	2.34E-03	0.00E+00	5.07E-04	8.09E-04	1.24E-03	1.87E-02
NumOfSportClubs	1.59E-04	2.85E-04	0.00E+00	0.00E+00	0.00E+00	1.73E-04	1.26E-03

Figure 3 shows the distributions and regression plots of all the features. Looking at the distributions of the variables, all age intervals seems to normally distributed. The other variables have seem to have two or three underlying distribution, which could be helpful for clustering the data.
None of the variable combinations seem to have enough correlation to make one of data features redundant. The most correlation seems to present between neighbouring age interval, which can be expected. Also, visually no obvious clusters emerge when plotting two variables against each other. The neighbourhood Feijenoord also do not seem to have one distinct variable different from the neighbourhoods in Amsterdam.

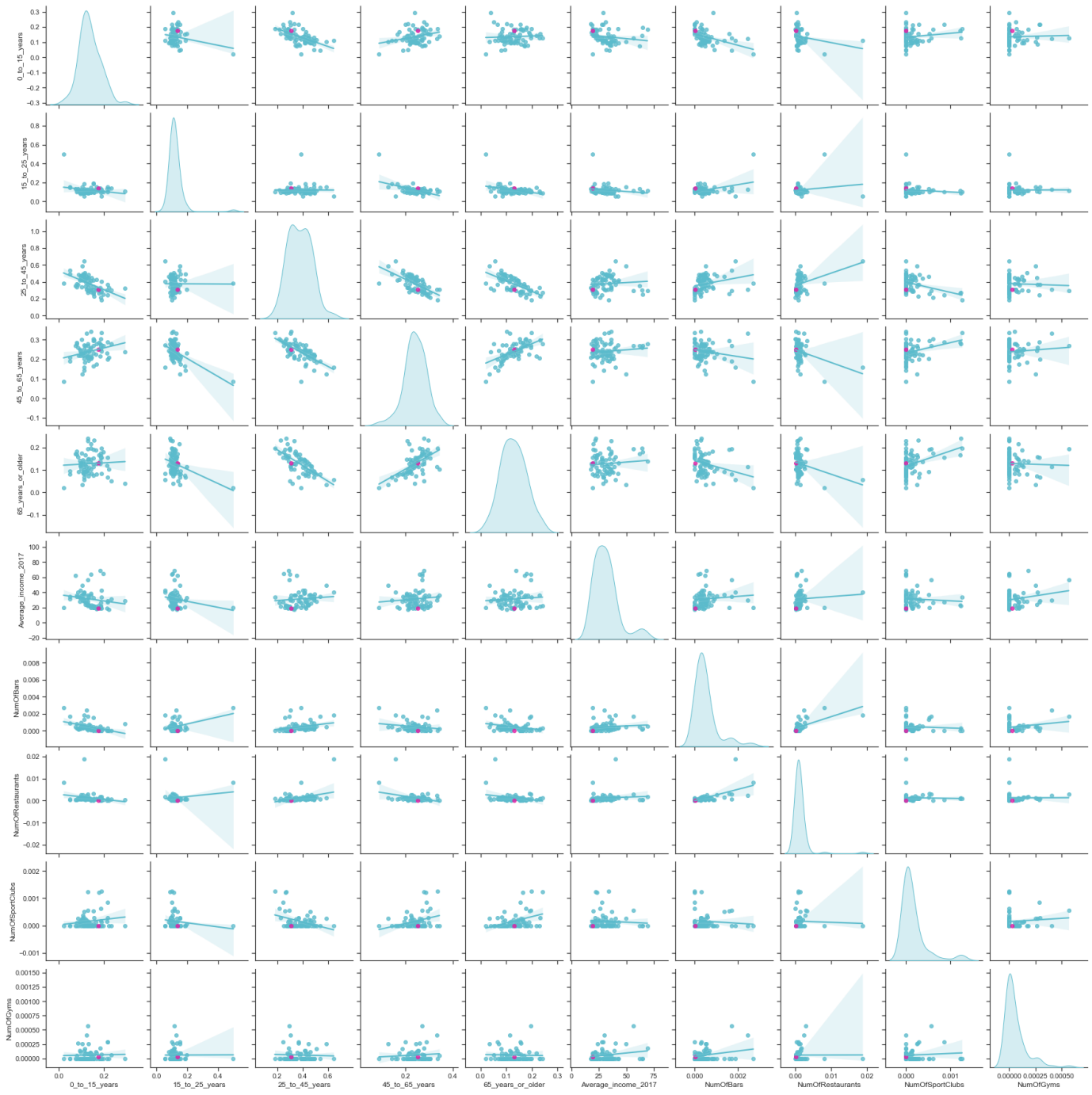

Figure 3: Distributions and regression plots of the features for each borough (purple dot = Feijenoord).

3.2 Modelling

The data is unlabelled and the question is which boroughs (observations) are similar to Feijenoord Rotterdam. This means that the data has to be clustered. This can be done with the k-means algorithm. First, the optimal number of clusters needs to be determined. Two methods have been compared to find the optimal number of clusters: the elbow method and the silhouette method. Using the elbow method the number of clusters is plotted against the sum of squared distances. The optimal number of clusters will be at the 'elbow' of the plot where the sum of squared distances does not decreases as rapidly with increasing clusters as before. The left graph in figure 4 shows the elbow plot for the k-means algorithm applied on this data. No clear elbow point could be distinguished. The optimal number of clusters varies between the 6 and 8.

Because the optimum was not clear using this method alone, the silhouette method was also used. In this method the silhouette score is plotted for the range of k's. This score measures how well a datapoint lies within a cluster, with one being perfect and 0 being close to the boundaries. A negative score means the point is outside the boundary. The right figure shows that the maximum silhouette score is at 6 clusters, matching the results of the elbow method. Therefore, 6 clusters was chosen as optimal.

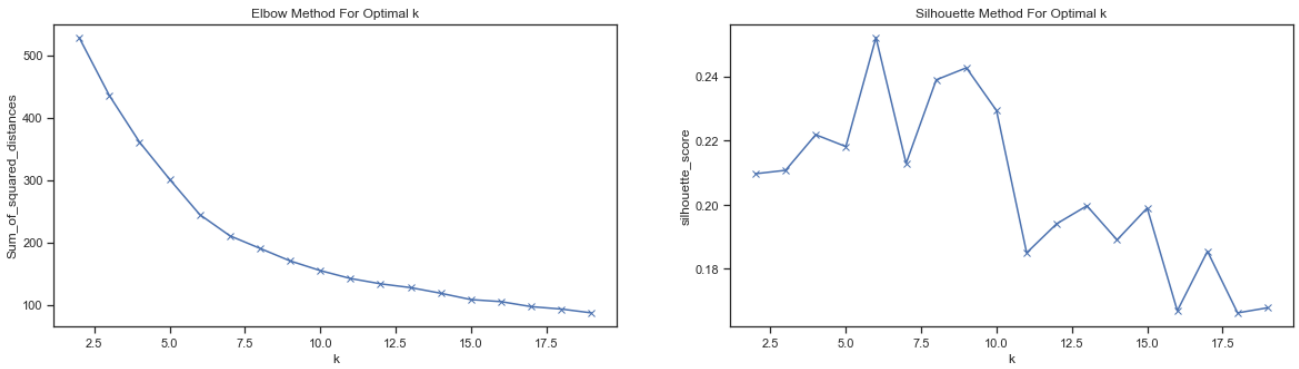

Figure 4: The optimal number of clusters with the elbow method (left) and the silhouette method (right)

4. RESULTS

4.1 Visualizing the clusters on the map

The clustering of the boroughs can be used to identify the areas similar to that of Feijenoord, Rotterdam. Boroughs that are close together geographically also tend to group together in one cluster. The coordinates of the boroughs has not been included for modelling the data. It can be expected that neighbouring boroughs are also similar in population distribution and venues. Although no conclusions can be drawn from these conclusion, it is indicative that the clustering of the data went as expected. Furthermore, it can be seen that the boroughs falling in the same cluster as Feijenoord are mostly in the suburbs, similar to Feijenoord. However, from this map no conclusion can be drawn about the basis upon which the clustered are created.

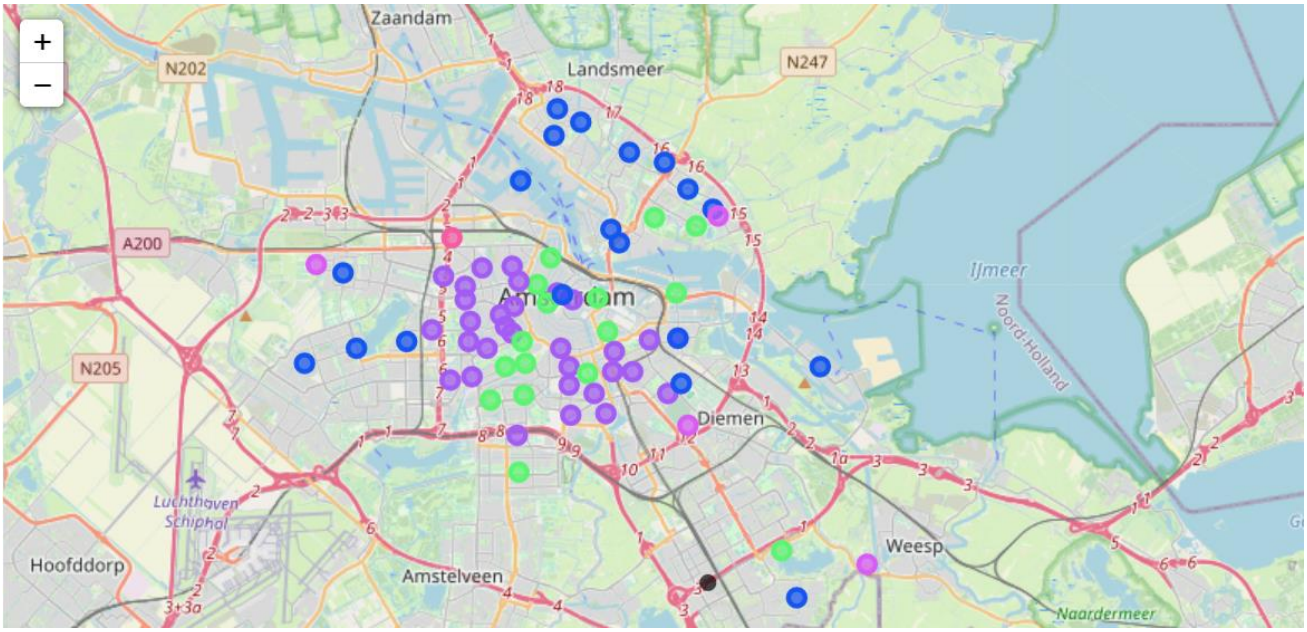

Figure 5: Results of clustering, with the colours denoting the 6 clusters. The blue colour denotes boroughs in the same cluster as Feijenoord.

4.2 Zooming in on the clusters

Figure 6 shows the scatterplots of the different variables and the densities of the variable, all grouped per cluster indicated with the colours. Not one variable (combination) shows a true separation between the different clusters. The most interesting distribution is that of the average income of the boroughs. For the cluster that contains Feijenoord the average income seems to be low compared to the other clusters. On average the number of venues seems to be lower for the cluster of Feijenoord, while the number of Sport Clubs seem to more average. This can also be expected as in the suburbs, in general more space is reserved for sport clubs. The age distribution for the Feijenoord cluster seems to be somewhat lower to average compared to the other boroughs.

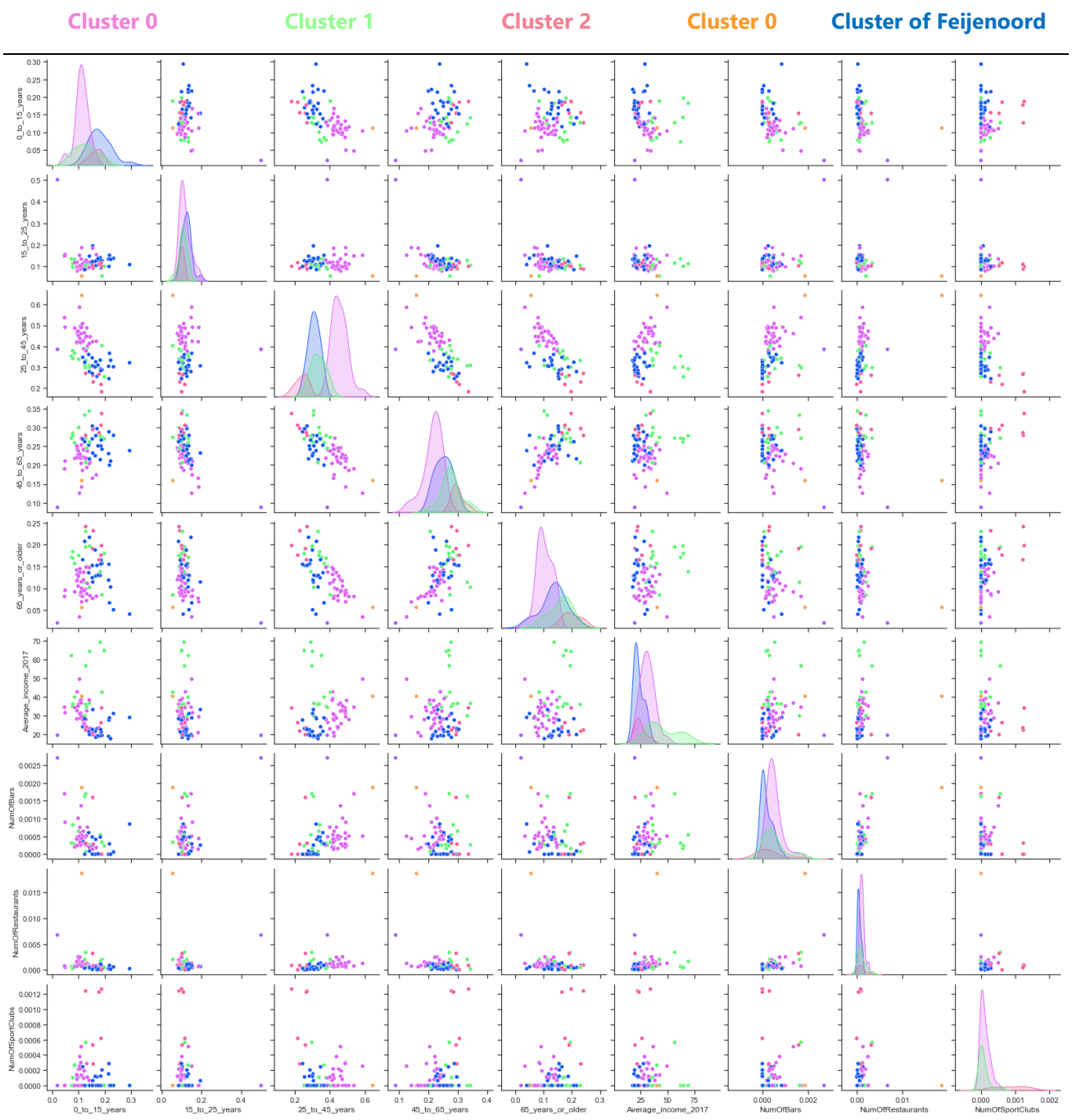

Figure 6: Distributions and regression plots for the variables used in clustering the boroughs. The colours denote the clusters. The blue colour denotes the cluster to which Feijenoord belongs.

4.3 Competition

Now that the similar boroughs to Feijenoord, Rotterdam are known for Amsterdam, a comparison on the competition for the gym in these areas. With the Foursquare data we also retrieved the number of gyms in a given area. This makes the comparison straightforward. However, not all boroughs are similar in size or population. Therefore, the number of gyms has been normalized to the total population in the area, in the form of the "number of gyms/ 100.000 persons". The boroughs in the cluster of Feijenoord have been sorted based on these values and the resulting table is shown in table 3.

The table shows that there are 11 boroughs that are similar to Feijenoord, and will not lead to extra competition in that area. Furthermore, there are 8 boroughs that are similar to Feijenoord, but do already have gyms present in that area, making them less suited as a new location for the gym.

Name	0 -15 year	15 – 25 year	25-45 year	45 – 65 year	65< year	Average income 2017	# Of Bars	# Of Restaurants	# Of Sport Clubs	# Of Gyms/ 100.000 persons
Houthavens	0.19	0.06	0.40	0.27	0.08	43	0.00034	0.0020	0.00000	0
Slotervaart Noord	0.18	0.13	0.28	0.25	0.16	23	0.00000	0.0011	0.00012	0
De Punt	0.18	0.14	0.33	0.21	0.13	19	0.00048	0.0011	0.00000	0
IJburg Zuid	0.29	0.11	0.32	0.24	0.04	29	0.00085	0.0002	0.00000	0
Geuzenveld	0.22	0.13	0.33	0.21	0.10	18	0.00000	0.0001	0.00000	0
Volewijck	0.17	0.10	0.33	0.28	0.13	18	0.00042	0.0006	0.00000	0
IJplein/Vogelbuurt	0.15	0.11	0.34	0.27	0.12	20	0.00049	0.0005	0.00000	0
Tuindorp Oostzaan	0.19	0.09	0.29	0.27	0.16	21	0.00000	0.0002	0.00000	0
Oostzanerwerf	0.17	0.13	0.25	0.30	0.15	23	0.00000	0.0000	0.00011	0
Gein	0.15	0.12	0.27	0.30	0.15	23	0.00000	0.0001	0.00000	0
Waterlandpleinbuurt	0.21	0.13	0.30	0.23	0.12	19	0.00000	0.0002	0.00015	0
Feijenoord	0.17	0.14	0.31	0.25	0.13	19	0.00001	0.0001	0.00000	3
Middenmeer	0.15	0.20	0.31	0.23	0.11	33	0.00025	0.0006	0.00006	6
IJburg West	0.23	0.14	0.30	0.28	0.05	31	0.00006	0.0006	0.00000	6
Banne Buiksloot	0.19	0.13	0.29	0.24	0.16	19	0.00000	0.0001	0.00020	7
Buikslotermeer	0.14	0.15	0.28	0.22	0.22	20	0.00009	0.0004	0.00000	9
Indische Buurt Oost	0.14	0.11	0.36	0.25	0.14	22	0.00061	0.0009	0.00000	10
Osdorp-Oost	0.16	0.12	0.30	0.22	0.19	22	0.00000	0.0005	0.00012	12
Kadoelen	0.22	0.10	0.24	0.29	0.15	28	0.00000	0.0003	0.00029	29
Elzenhagen	0.22	0.15	0.37	0.20	0.07	29	0.00029	0.0009	0.00000	29

5. CONCLUSION

The clustering of the data showed that is not one variable that can be used to cluster the boroughs. Instead, the combination of features is necessary. The goal of this project was to find areas in Amsterdam, similar to the Feijenoord, Rotterdam. Data of the CBS giving the age distributions and average income in the boroughs and location data of Foursquare giving the number venues in the area have been used to cluster the boroughs in Amsterdam and the borough Feijenoord. Clustering with k-Means showed that the boroughs close together geographically, also tended to cluster together with the k-means clustering without including the location in the model. This gave confidence that the clustering went as expected.

The cluster to which Feijenoord belong, were positioned mostly on the outskirts of the city, similar Feijenoord in Rotterdam. When looking at the distributions of the variables between clusters, the most relevant observation was that the cluster of Feijenoord had a relative low average income compared to the other clusters. Also the age seemed to be low to average.

The boroughs were filters on their respective clusters, only including the cluster of Feyenoord. Based on the number of gyms/100.000 persons the boroughs were sorted. This resulted in the boroughs in Amsterdam most similar to Feijenoord with little competition. The selected boroughs were:

- Houthavens
- Slotervaart Noord
- De Punt
- IJburg Zuid
- Geuzenveld
- Volewijck
- IJplein/Vogelbuurt
- Tuindorp Oostzaan
- Oostzanerwerf
- Gein
- Waterlandpleinbuurt

REFERENCES

[1] https://opendata.cbs.nl/statline/portal.html, date accessed: 02/04/2021

[2] https://foursquare.com/, date accessed: 02/04/2021