Who needs customers anyway?

Finding the best place to build a cinema: Coursera IBM Data Science Capstone Project

Introduction

There are many cinemas in the Netherlands. The last decennium has seen a hectic business landscape in which many newer modern multiplexes were built, and older historic smaller cinemas have found it difficult to survive. Whenever you build a new place where people can have fun, they will come. But not always in large numbers. Some municipalities have many potential customers and others have few. If the wrong one is chosen, the new cinema will fail, costing a lot of money.

The question is: which municipalities would be the best to build your new cinema?

This project uses data to determine these places and is therefore very useful for entertainment companies that wish to expand with a cinema in the Netherlands. Also, companies with the wish to build something substitutional with cinemas such as theaters or arcades will find this analysis interesting. Last, the municipalities themselves can use the analysis to determine how viable the construction of a cinema in their municipality is and if the necessary infrastructure for that is a good idea.

Data

To assess the best place to build a cinema, I will focus on the municipality with the highest number of potential customers. For this, I need the following data:

- The current number of cinemas per municipality --> Foursquare
- The longitude and latitude of each municipality --> geopy
- Information for all the municipalities in the Netherlands, such as number of inhabitants and land area --> Wikipedia

The current number of cinemas and the number of inhabitants will give an indication of the number of potential customers and are the most important.

The longitude and latitude is used to be able to extract the information from Foursquare.

The other information for all the municipalities, along with the longitude and latitude is used in the machine learning to find any patterns.

Methodology

For the initial exploration of the data, the municipal data is plotted vs longitude and latitude, and several parameters are plotted against each other. Linear regression is performed to see if there is a correlation between a parameter and the number of cinemas.

Next, k-means clustering is performed on some of the parameters, where parameters that are too similar are removed to prevent over-weighing them, e.g. Land area and Total area. To determine the optimal number of clusters chosen, both the elbow method of the within-cluster sum of squares and the highest silhouette score is investigated. K-means clustering will give a label to each municipality.

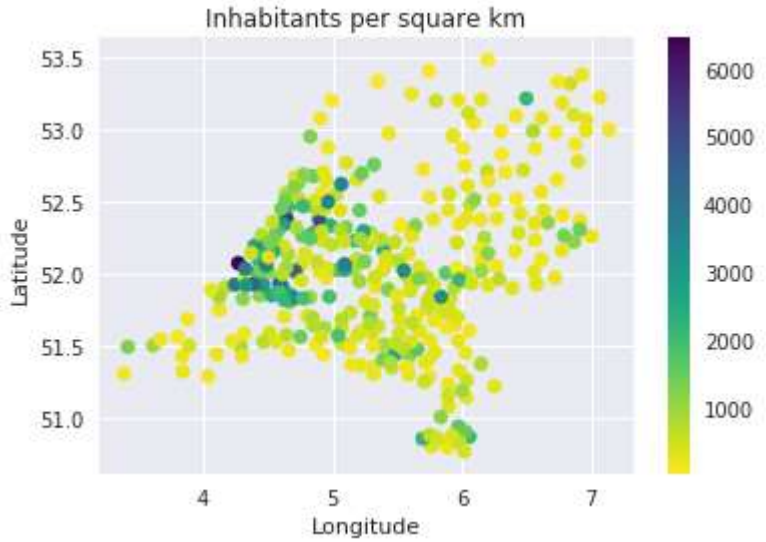
The groups with the labels arising from the clustering are investigated spatially and in histograms, to see if there are patterns that can be used for the search of the best cinema location.

Spatial explorative analysis

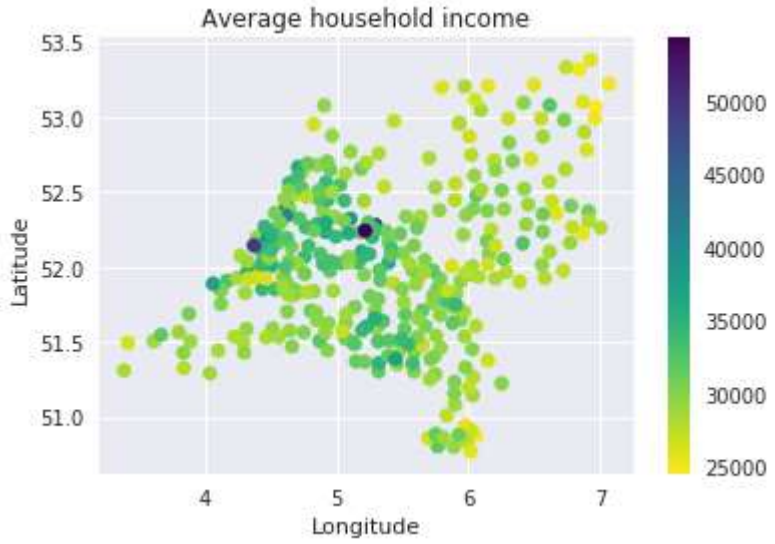
The Netherlands consist of twelve provinces. The Randstad is an unofficial area in the west including Amsterdam, Rotterdam, The Hague, and Utrecht.

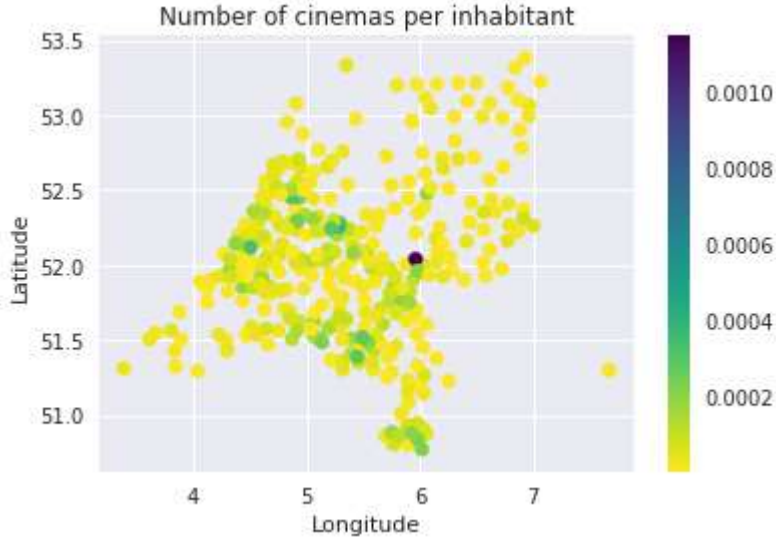
(image from Wikipedia)

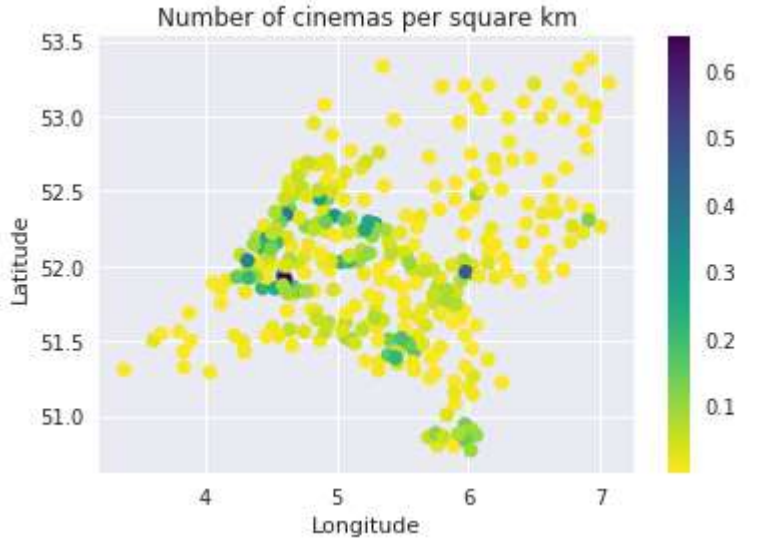
Exploring the initial data spatially shows that the Randstad has the highest population density.



The average household income and the number of cinemas per inhabitant and per square km are more evenly distributed, but have a focal point towards the Randstad, Noord-Brabant, and Gelderland.





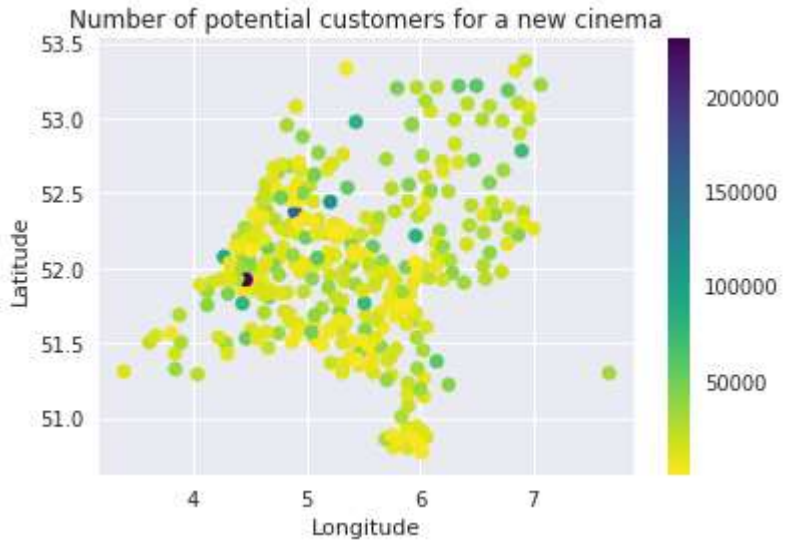


The potential number of customers for a new cinema is spread more evenly across the provinces of the Netherlands, but with local fluctuations within provinces.

This parameter 'Number of potential customers for a new cinema' is viewed as follows:

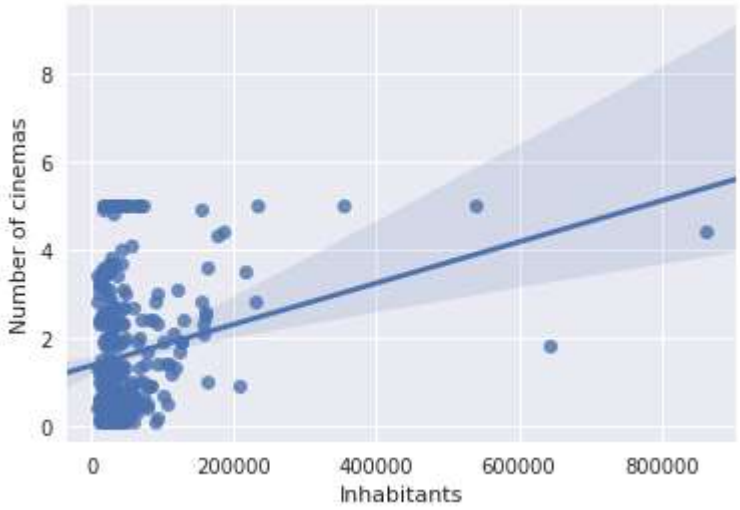
To make life simple, it is assumed that people only go to one cinema and in their own municipality. Also, a new cinema doesn't lead to more potential customers in the municipality.

If a municipality has X inhabitants and Y cinemas, each cinema has X/Y customers. Now, a new cinema will potentially have X/(Y+1) customers.

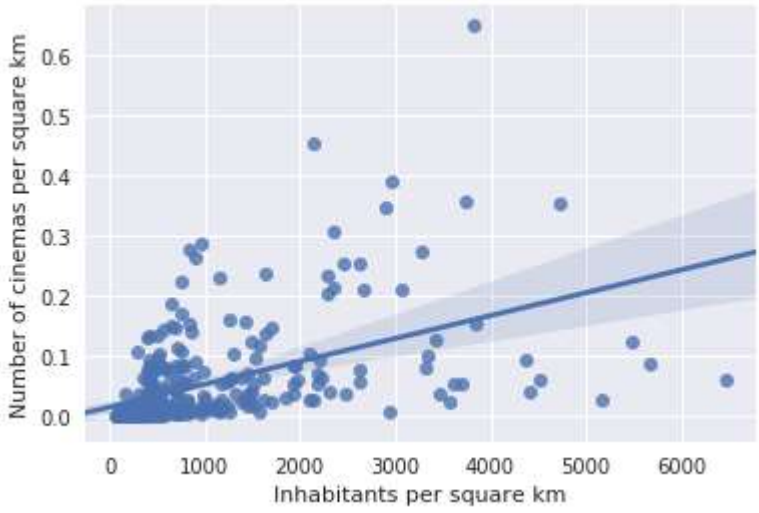


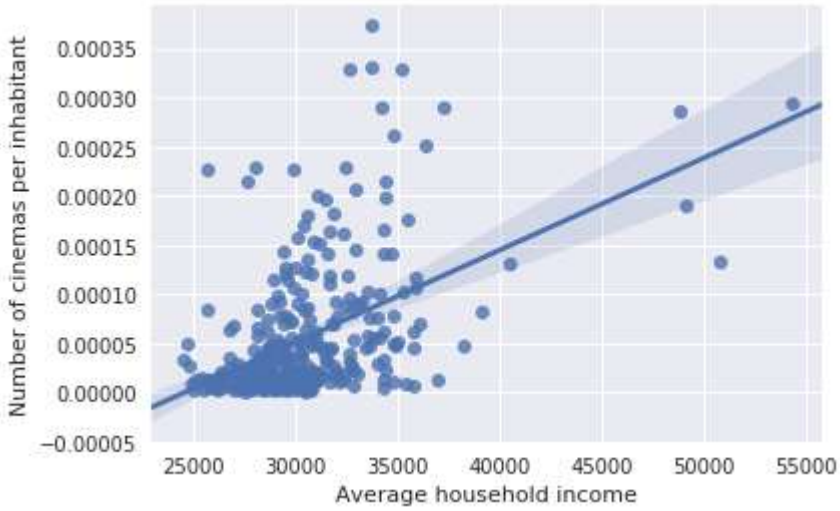
Correlation explorative analysis using linear regression

Some correlations are expected, such as: municipalities with more inhabitants have more cinemas. Using linear regression, this turns out to have a poor correlation actually (R squared is only 0.06).



Inhabitants per square km with Number of cinemas per square km (R squared = 0.25) and Average household income with Number of cinemas per inhabitant (R squared = 0.23) have a slightly better correlation, but it is still quite poor.





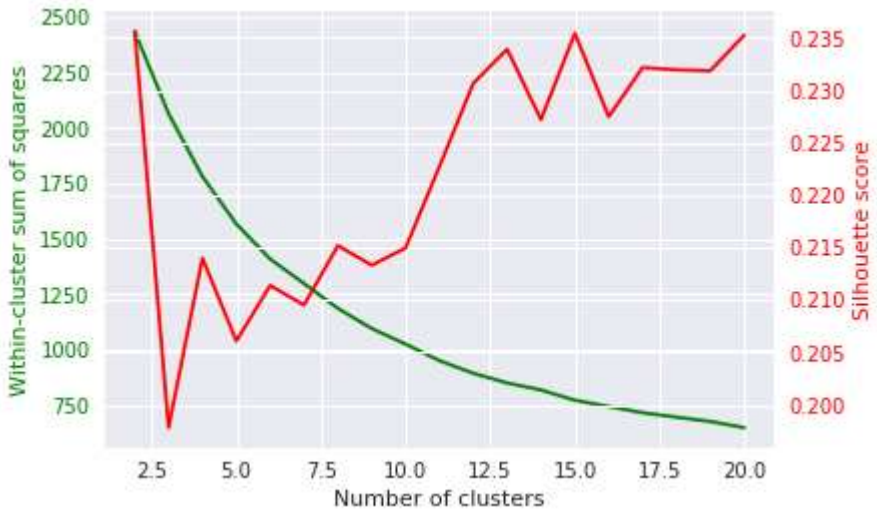
Machine learning: K-means clustering

With all these parameters for the municipalities, we can wonder if some municipalities are similar to others. This can be done with the unsupervised machine learning method of K-means clustering. This method tries to find clusters of data points (municipalities) that have similar parameter values, while being different from data points in other clusters.

When using K-means clustering, it is important to use the most suitable number of clusters. I used two methods to determine this. Both involve running the clustering and the recording a figure of merit: either the within-cluster sum of squares or the silhouette score. The elbow method of the within-cluster sum of squares takes as the optimal number of clusters the one where the parameters in the suddenly/sharply look less similar. The silhouette method takes as the optimum number of clusters where the one where the data points are optimally far from other clusters.

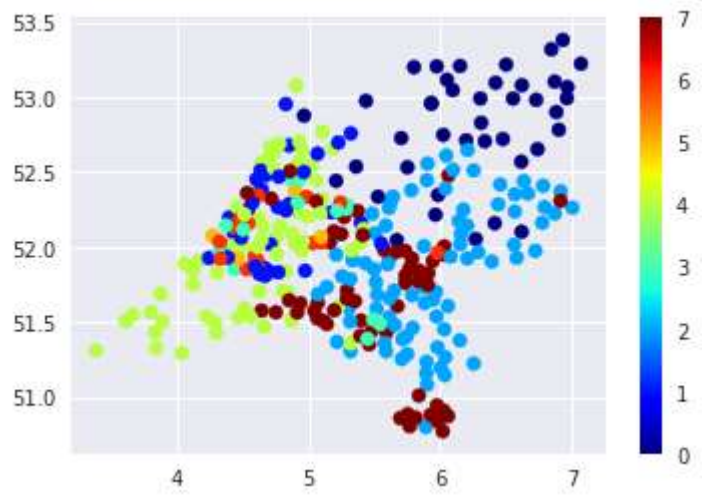
The figure below shows that it is very difficult to find an elbow (or sharp decrease) in the within-cluster sum of squares (green graph). I chose not to use this method for that reason.

The silhouette score is very high for two clusters. This makes sense as the cluster centers are far from each other. At 12 or more clusters, the silhouette score also becomes high. This is because the silhouette score also takes into account the size of the cluster a data point is in. At so many clusters, this becomes very small. Both regions are therefore not representative. I chose the number of clusters with the highest silhouette score in the intermediate region, being eight clusters.



K-means clustering gave a clear distribution of the clusters, both geographical (visible in the graph below) and for the average value of the other parameters (see table):

0. north, few cinemas, lower average household income
1. Randstad, some cinemas, lower average household income
2. east, few cinemas, lower average household income
3. scattered, many cinemas, high average household income
4. Randstad, few cinemas, lower average household income
5. big 4 cities, many cinemas, lower average household income
6. Randstad, many cinemas, above average average household income
7. center/south/east, many cinemas, lower average household income



Label	Inhabitants	Land area	Inhabitants per square km	Average household income	Latitude	Longitude	Number of cinemas
0	55968	221	328	28000	52.81	6.25	0.6
1	67107	28	2596	29465	52.22	4.76	1.6
2	38206	97	453	29389	51.85	5.94	0.7

3	16132	28	643	39013	52.06	4.96	3.9
4	34688	87	525	31880	52.00	4.61	0.9
5	599628	140	4565	28550	52.11	4.68	4.1
6	39248	14	2861	31940	52.06	4.76	4.0
7	48224	52	898	30343	51.64	5.51	3.1

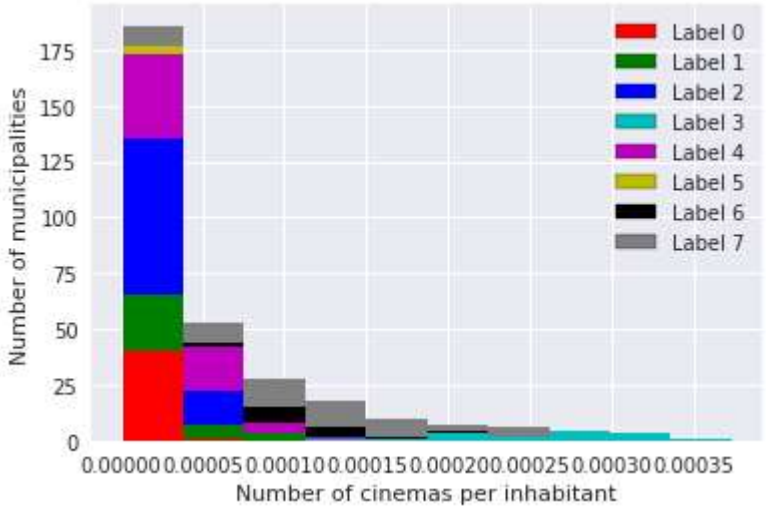
Determination of metric for best municipality

Four parameters were calculated, which could be used as a metric to determine the best municipality to build a new cinema:

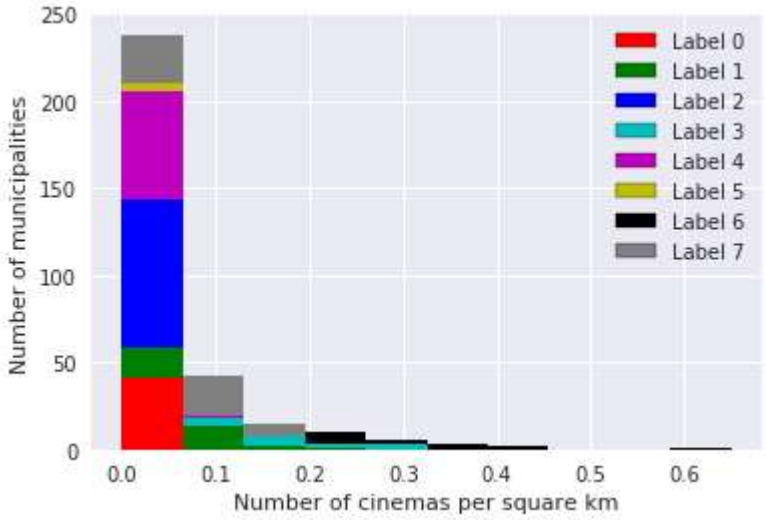
1. Number of cinemas per inhabitant
2. Number of cinemas per square km
3. Number of cinemas per inhabitant times square km
4. Number of potential customers for a new cinema

I use qualitative argumentation and look at histograms of the labels with the parameter under investigation to determine the best metric.

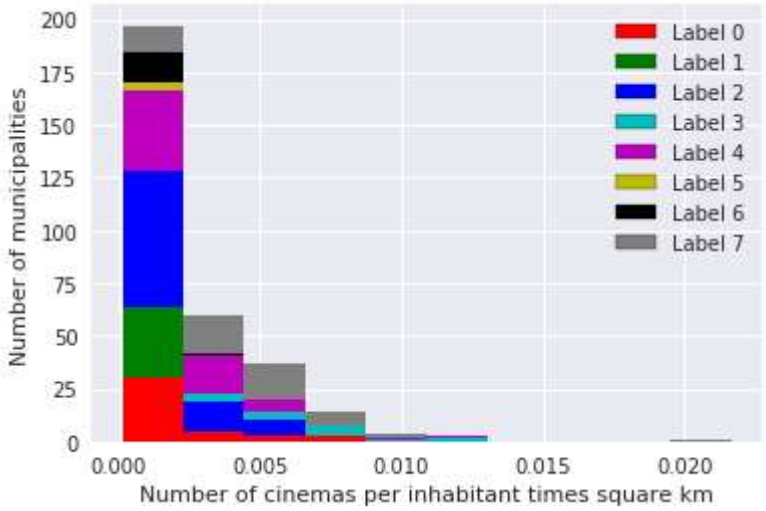
For a niche, you would want the Number of cinemas per inhabitant to be as low as possible. If there are fewer cinemas now, they could use one. But, would you also get the most customers? This is not addressed by using this parameter as a metric.



For a niche, you would also want the Number of cinemas per square km to be as low as possible. If people have to travel far to reach a cinema, they could use one. But perhaps this is a sparsely populated area, and few potential customers live there?

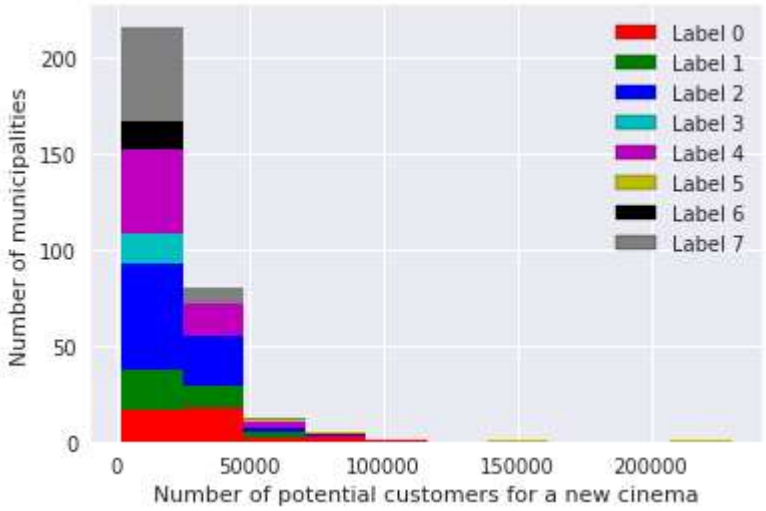


Looking for densely populated areas with few cinemas gives a decent indication for the location of the next cinema. But using the Number of cinemas per inhabitant times square km doesn't guarantee if these municipalities have the largest number of potential customers.



In the histograms above, we can also see that these criteria give a lot of municipalities that are eligible. This makes it difficult to choose.

Simpler would be to look at the potential number of customers more directly. We could look per municipality at the number of inhabitants divided by the current number of cinemas + 1 (the new cinema). This also gives fewer municipalities to choose from.



I choose the Number of potential customers for a new cinema as the metric to determine the best municipality to build a new cinema.

Results

According to this analysis, the 10 best municipalities to build a new cinema are:

1. Rotterdam
2. Amsterdam
3. Almere
4. Den Haag
5. Súdwest-Fryslân
6. Apeldoorn
7. Oss
8. Emmen
9. Groningen
10. Venlo

Municipality	Number of potential customers for a new cinema	Inhabitants	Land area	Average household income	Number of cinemas	Label	Province
Rotterdam	230,188	644,527	218	26,200	1.8	5	Zuid-Holland
Amsterdam	159,852	863,202	166	29,700	4.4	5	Noord-Holland
Almere	109,378	207,819	129	28,800	0.9	0	Flevoland
Den Haag	89,664	537,988	82	28,300	5	5	Zuid-Holland
Súdwest-Fryslân	81,550	89,705	522	27,500	0.1	0	Friesland
Apeldoorn	81,228	162,456	340	29,200	1	0	Gelderland
Oss	76,197	91,437	163	29,100	0.2	2	Noord-Brabant
Emmen	71,407	107,111	335	26,200	0.5	0	Drenthe
Groningen (gemeente)	60,882	231,354	95	25,700	2.8	0	Groningen (provincie)
Venlo	59,751	101,578	124	26,800	0.7	2	Limburg

Discussion

The top 10 of best municipalities seems to consist of two groups:

1. Large cities (Rotterdam, Amsterdam, Den Haag) which number of cinemas is suspiciously low (Rotterdam), low (Amsterdam), or capped by Foursquare (Den Haag)
2. Medium-sized cities with relatively few cinemas

One extra cinema in these municipalities can expect a lot of customers either because there are many people (group 1) or there are relatively few cinemas to compete with (group 2).

The top 10 only has labels 0, 5, and 2. The top 50 does not have labels 3 and 6, which are both smaller municipalities with relatively many cinemas. It is worth mentioning that the top 10 features municipalities from 9 different provinces, only Zuid-Holland is featured twice.

The top 10 has an average household income that is lower than the average of the Netherlands. This analysis does not take into account the behaviour of consumers. Do people with more disposable income spend more or less money going to the cinema? This is a serious shortcoming of this analysis and is recommended to be taken into account for following studies.

The usefulness of this result is seriously diminished by the low reliability of the Foursquare data. It goes beyond the scope of this course to improve that. Also, more reliable analysis could be done with data on a finer level than on municipal level only.

Conclusion

The top 10 of best municipalities to build a new cinema were found to be:

1. Rotterdam
2. Amsterdam
3. Almere
4. Den Haag
5. Súdwest-Fryslân
6. Apeldoorn
7. Oss
8. Emmen
9. Groningen
10. Venlo

This result was obtained using the number of potential customers for a new cinema.

K-means clustering was performed with eight clusters, of which only three clusters are featured in the top 10.