IBM Data Science

Capstone Project Report

Tennyson Wu

16 November 2020

# Introduction

## Background

In an increasingly globalised world, more and more individuals are finding opportunities to work abroad. Much can be found online about the new city that individuals may be moving to. However, it may not always be so easy to find the right neighbourhood to move to. The aim of this project is to compare the neighbourhoods in two cities to determine how similar the cities are to live in, and to assist the individual in finding a suitable neighbourhood in the destination city.

Personally, I spent a year in Amsterdam while completing my master’s program. It was a great experience, so much so that I am currently searching for work in the Netherlands. Some incredibly interesting firms are in Rotterdam. Thus, the question for me is, how similar are the neighbourhoods in Rotterdam to the ones in Amsterdam.

According to Geonames.org (Geonames, 2020) Rotterdam is the second largest city in the Netherlands, second only to Amsterdam. Yet, Rotterdam is known as a port city, having the largest port in Europe (Holland.com, 2020) while Amsterdam is widely known as a famous tourist destination.

Therefore, these two cities have fundamentally different backgrounds, despite both being large cities in the Netherlands.

This project will attempt to apply k-means clustering to the neighbourhoods in Amsterdam and Rotterdam to better understand these two cities.

## Research Questions

To specify the aims of this project more clearly, the following questions will be answered:

How similar is Rotterdam to Amsterdam?

Would I enjoy living in Rotterdam?

Based on my experience in Amsterdam neighbourhoods, which neighbourhoods in Rotterdam would I enjoy living in?

## Interest

This type of research would likely be most helpful to individuals who are moving cities, where they could take the process used here and apply it to their own specific locations. For example, they might be living in one city in the world and may be moving to another, somewhere across the globe.

Additionally, companies located in Rotterdam may be interested in the results of this specific project to better understand the similarities between the famous Amsterdam and their own city. These insights could help them better attract international talent that have started work in Amsterdam.

Similarly, the city of Rotterdam itself may also use the results of this study, should they wish to promote their city to residents of Amsterdam.

# Data

## Data Sources

To compare the neighbourhoods in Amsterdam and Rotterdam, two sets of data are required. Lists of neighbourhoods for Amsterdam and Rotterdam, along with their geographic coordinates are needed to identify the neighbourhoods and place them on a map. The venues in each neighbourhood, what type of venues they are, and their geographic coordinates are also needed to determine differences and similarities between neighbourhoods.

Lists of neighbourhoods for Amsterdam and Rotterdam were found on Wikipedia (Wikipedia, 2020). These lists showed 105 and 90 neighbourhoods for Amsterdam and Rotterdam, respectively. Most of the geographic coordinates for these neighbourhoods were found through an online platform (Coordinates Finder, 2020), while the remainder were found through manual Google searches.

Foursquare is an online platform where individual users identify information about venues around the world (Foursquare, 2020). The types of venue data described above can be extracted from Foursquare through requests sent to the Foursquare API via Python code.

## Data Pre-processing

Geographic coordinates for the different neighbourhoods were checked for outliers through visual inspection of the coordinates data. Given these neighbourhoods were in close proximity to each other, any coordinates that appeared to deviate too far from the norm were checked manually through Google searches and corrected where needed.

Maps were then drawn for the neighbourhoods in Amsterdam and Rotterdam. Figure 1 shows a map of Amsterdam, along with all its neighbourhoods indicated as blue dots.

|  |
| --- |
| **Figure 1**  **Map of neighbourhoods in Amsterdam** |
|  |

A visual inspection of the Rotterdam map showed an additional outlier that was not found in the previous process. For illustration purposes this map of Rotterdam is shown in Figure 2, along with all the neighbourhoods of Rotterdam, indicated as red dots.

|  |
| --- |
| **Figure 2**  **Map of neighbourhoods in Rotterdam** |
|  |

The neighbourhood with coordinates in the city of Delft to the north of Rotterdam is labelled as Nooddorp. This neighbourhood was on the list of Rotterdam neighbourhoods according to Wikipedia (Wikipedia, 2020). However, further investigation revealed that Nooddorp was a term for neighbourhoods around Rotterdam where emergency shelters were set up during World War 2 (Wikipedia, 2020). Despite best efforts, no neighbourhood named Nooddorp in present day Rotterdam could be found on a map., thus it was dropped from further analysis.

After the venue data was obtained via the Foursquare API, a sample of the data set was displayed to confirm the correct types of data loaded. Table 1 shows a sample of the data obtained from Foursquare.

|  |
| --- |
| **Table 1**  **Sample of venue data from Foursquare** |
|  |

This data was cross referenced with the lists of neighbourhoods to determine if data was found for all the neighbourhoods. No venue data was loaded for two neighbourhoods, Landelijk Noord in Amsterdam, and Vondelingenplaat in Rotterdam.

Since venue data on Foursquare is created by individual users who visit those locations it is entirely possible that the data is simply incomplete when compared to the real world. To confirm that venue data for these neighbourhoods were not simply failing to load, manual queries were performed on the Foursquare website. No relevant venue data was returned when querying Landelijk Noord; a few venues around Amsterdam Central were returned. No search results were found for Vondelingenplaat at all. Since there were only two neighbourhoods out of 195 with no data, these two neighbourhoods were excluded from further analysis.

## Data Preparation

To prepare the venue data for k-means clustering, the individual entries for each venue were grouped by neighbourhood and summarised by the type of venue. The total number of venues were then calculated for each neighbourhood.

For prior exercises using k-means clustering on neighbourhoods, the neighbourhoods were clustered by the top 10 most common venue types in each neighbourhood.

In this project, the totals for each neighbourhood were checked to determine how many neighbourhoods have very little venue data; it makes no sense to consider the top 10 most common venue types for a neighbourhood if it had less than 10 venues recorded. Out of the 192 neighbourhoods remaining in the study, 56 have fewer than 10 venues recorded on Foursquare.

56 neighbourhoods are too many to simply leave out of the study, for any results to meaningfully represent the two cities. To determine the next best course of action, the number of neighbourhoods with fewer than 5 venues were counted. There were 27 neighbourhoods with fewer than 5 venues listed on Foursquare. A similar count was performed for neighbourhoods with fewer than 3 venues listed, and 7 neighbourhoods were found.

Overall, the issue becomes a trade-off between having sufficient data for the results to represent the cities adequately and having enough complexity in features for k-means clustering to create good groupings.

After some consideration, the decision was made to perform analysis on both data for the top 5 most common venue types and the top 3 most common venue types in each neighbourhood. The expectation is that clusters based on the top 3 most common venue types will be less nuanced and possibly less helpful in describing similarities between neighbourhoods. It will be interesting to see what the actual results will be.

Two data sets were subsequently created. One dataset contained neighbourhoods with 5 or more venues listed and the top 5 most common venue type for each neighbourhood. The other contained neighbourhoods with 3 or more venues listed and the top 3 most common venue types for each neighbourhood.

# Methodology

The next decision was how many clusters would be optimal. There are several methods to determine the optimal value for k, the number of clusters (Datanovia, 2020). The methods applied in this case were the Elbow Method, the Silhouette Score and Davies Bouldin Index. For all three methods, k-means clustering was performed for various values of k, ranging from 2 to 20.

As discussed in the end of section 2.3, k-means clustering was performed twice. Once for the dataset focusing on the top 5 most common venue types and again for the dataset focusing on the top 3 most common venue types.

## Top 5 most common venue types

For the Elbow Method, the sum of square distances from each point to its assigned cluster centre were calculated, for each value of k, and then plotted on a graph. Graph 1 shows the results of these calculations.

|  |
| --- |
| **Graph 1**  **The Elbow Method** |
|  |

According to the Elbow Method, the optimal value for k would then be the “elbow” point, where the sum of square distances begin to decrease linearly (Geeks for Geeks, 2019). Inspection of the graph does not lead to a strong indication of the “elbow” point. It appears in this case k might be 5. Perhaps the other methods to determining k will assist in making a clearer decision.

Next, the Silhouette Score is a measure of how similar a neighbourhood is to its own cluster compared to other clusters, where higher scores indicate better clustering (Mahendru, 2019). Graph 2 shows the Silhouette Scores for different values of k.

|  |
| --- |
| **Graph 2**  **Silhouette Score** |
|  |

Based on the graph, the Silhouette Score is highest for two clusters. However, the Silhouette Scores appear to be quite low even when k is 2; scores range between -1 and 1 (Mahendru, 2019) so 0.17 is not a very high score. Additionally, splitting the neighbourhoods of Amsterdam and Rotterdam into only two clusters does not seem productive in terms of answering the questions in section 1.2. If all the neighbourhoods are split into two categories, it would likely remain difficult to determine which neighbourhood would be most suitable to move to.

Finally, the Davies Bouldin Index measures how distinct the clusters are from each other, where lower scores indicate more successful clustering. Graph 3 shows the Davies Bouldin Scores calculated for the different values of k.

|  |
| --- |
| **Graph 3**  **Davies Bouldin Score** |
|  |

The graph shows a local minimum score around where k is 5 or 6. Otherwise the scores appear to generally be a decreasing function of k and for the range displayed in the graph, the absolute minimum score is where k is 19.

Based on the heuristic for the Davies Bouldin Index, the optimal value of k would be 19. However, this is not in line with the results of the Elbow Method or the Silhouette Score. Unfortunately, all three methods recommend a different value for k, bringing into doubt the effectiveness of any k-means clustering performed. However, for the purposes of this open and exploratory project, the value for k to be used with the top 5 most common venue types will be 5.

## Top 3 most common venue types

Here the optimal number of clusters was sought through the same three methods as those used for the top 5 most common venue types. Graphs 4, 5, and 6 show the line graphs of the squared error, Silhouette Score, and Davies Bouldin Scores for different values of k, respectively.

|  |  |  |
| --- | --- | --- |
| **Graph 4** | **Graph 5** | **Graph 6** |
| **Elbow Method** | **Silhouette Score** | **Davies Bouldin Index** |
|  |  |  |

Similar to the data set with the top 5 most common venue types, these three measures provide conflicting recommendations for the optimal number of clusters, k. Again, the Elbow Method appears to weakly recommend 5 clusters, though there is an irregular dip in squared error for 10 clusters. The Silhouette Score suggests 2 clusters again as well. The Davies Bouldin Score recommends 13 clusters.

Since there is no clear recommendation for the value of k, 5 clusters will be used for the data set with top 3 most common venue types, in line with the decision for the data set with the top 5 most common venue types. This will allow more clear and easy comparison between the choices of using only top 3 compared to using the top 5 most common venue types for each neighbourhood.

# Results

K-means clustering was then run to group the neighbourhoods using the top 5 and top 3 most common venue types in each neighbourhood, respectively. Results of the clustering process will be shown for the data using top 5 most common venue types first.

## Top 5 most common venue types

The clusters for the top 5 most common venue types are labelled 0 to 4 by default. Given there were over a hundred neighbourhoods, the full results are not presented in this report. For illustrative purposes, table 2 shows a sample of the results for cluster 0.

|  |
| --- |
| **Table 2**  **Sample of data for cluster 0 based on top 5 most common venue types** |
|  |

Based on inspection of the various clusters, the following conclusions were drawn for each cluster:

* Cluster 0 contains 45 neighbourhoods with the majority located in Rotterdam. For most neighbourhoods in this cluster the most common venue type is a supermarket.
* Cluster 1 contains 9 neighbourhoods with the majority located in Rotterdam. Bus stops are a common factor for all these neighbourhoods, being the most common venue type for most.
* Cluster 2 contains 3 neighbourhoods, two of which are in Amsterdam. The common venue type is soccer field, being the most common for two neighbourhoods and the third most common venue type for the last neighbourhood in this cluster.
* Cluster 3 contains 97 neighbourhoods, the majority of which are in Amsterdam. Bars, cafes, coffee shops and restaurants featured very prominently in the top 5 most common venue types for these neighbourhoods.
* Cluster 4 contains 11 neighbourhoods, the majority of which are in Amsterdam. This cluster is more varied than the previous ones, with a mix of outdoor venues such as parks and sports centres, as well as a variety of shops, supermarkets, and other venue types. The results of this cluster may indicate the possibility the clustering would have been better with 4 clusters instead of 5.

What follows are two maps, one for each city. The clusters on both maps are colour-coded such that a coloured dot on one map corresponds to the same cluster of neighbourhoods as dots of the same colour in the other map. Figure 3 shows a map of Amsterdam along with the clustering results of its neighbourhoods.

|  |
| --- |
| **Figure 3**  **Clusters of neighbourhoods in Amsterdam** |
|  |

Cluster 0 (supermarkets) is red, cluster 1 (public transport) is purple, cluster 2 (soccer fields) is blue, cluster 3 (bars, cafes, coffee shops) is green, and cluster 4 (mixed venues) is orange.

Very clearly, Amsterdam has many green neighbourhoods, filled with bars, cafes, and coffee shops. This is as expected of the world-famous tourist city.

Figure 4 shows a map of Rotterdam along with the clustering results of its neighbourhoods.

|  |
| --- |
| **Figure 4**  **Clusters of neighbourhoods in Rotterdam** |
|  |

Like Amsterdam, Rotterdam also has many green neighbourhoods in its centre, filled with bars, cafes, and coffee shops. However, it appears Rotterdam also has many red neighbourhoods where supermarkets are the most common venue type.

## Top 3 most common venue types

Overall, the results for the clustering of the top 3 most common venue types were similar to the results for the top 5 most common venue types, with a few exceptions. For illustrative purposes, table 3 shows a sample of the results for cluster 0.

|  |
| --- |
| **Table 3**  **Sample of data for cluster 0 based on top 3 most common venue types** |
|  |

Based on inspection of the various clusters, the following conclusions were drawn for each cluster:

* Cluster 0 contains 55 neighbourhoods with the majority located in Rotterdam. For most neighbourhoods in this cluster the most common venue type is a supermarket. Though it is a smaller majority compared to respective cluster formed from the top 5 most common venues.
* Cluster 1 contains 16 neighbourhoods with the majority located in Rotterdam. Bus stops are a common factor for most of these neighbourhoods, save for three. Again, this shows a slightly weaker link to bus stops compared to the respective cluster from the top 5 most common venue types.
* Cluster 2 contains 96 neighbourhoods, the majority of which are in Amsterdam. Bars, cafes, coffee shops and restaurants featured very prominently in the top 3 most common venue types for these neighbourhoods. Again, the cluster is not quite as clearly defined by this factor compared to the respective cluster from the top 5 most common venues.
* Cluster 3 contains 15 neighbourhoods, the majority of which are in Rotterdam. Surprisingly, this cluster has no comparative cluster from the top 5 most common venues. For the far majority of neighbourhoods in this cluster the most common venue type is tram station. Additionally, the only neighbourhood with a tram station in the top 3 most common venues has light rail stations as the third most common venue type.
* Cluster 4 contains 3 neighbourhoods, two of which are in Amsterdam. The common venue type is park, being the most common for two neighbourhoods and the third most common venue type for the last neighbourhood in this cluster. This is slightly different to the respective cluster from the top 5 most common venue type, where the common denominator was soccer field.

What follows are two maps, one for each city. The clusters on both maps are colour-coded such that one coloured dot on one map corresponds to the same cluster of neighbourhoods as dots of the same colour in the other map.

Figure 5 shows a map of Amsterdam along with the clustering results of its neighbourhoods.

|  |
| --- |
| **Figure 5**  **Clusters of neighbourhoods in Amsterdam based on top 3 most common venue types** |
|  |

Cluster 0 (supermarkets) is red, cluster 1 (bus stops) is purple, cluster 2 (bars, cafes, coffee shops) is blue, cluster 3 (tram stations) is green, and cluster 4 (parks) is orange.

Very clearly, Amsterdam has many blue neighbourhoods, filled with bars, cafes, and coffee shops. This is as expected of the world-famous tourist city.

Figure 6 shows a map of Rotterdam along with the clustering results of its neighbourhoods.

|  |
| --- |
| **Figure 6**  **Clusters of neighbourhoods in Rotterdam based on top 3 most common venue types** |
|  |

Like Amsterdam, Rotterdam also has many blue neighbourhoods in its centre, filled with bars, cafes, and coffee shops. However, it appears Rotterdam also has many red neighbourhoods where supermarkets are the most common venue type.

# Discussion

There were a few differences in the results of k-means clustering for top 5 and top 3 most common venue types per neighbourhood. However, the core findings appear to have strong similarities.

As mentioned in section 1.2, the questions to answer are:

1. How similar is Rotterdam to Amsterdam?
2. Would I enjoy living in Rotterdam?
3. Based on my experience in Amsterdam neighbourhoods, which neighbourhoods in Rotterdam would I enjoy living in?

After the analysis we can attempt to answer these questions.

## Similarities between Amsterdam and Rotterdam

Overall, the neighbourhoods in Rotterdam appear to share many similarities with Amsterdam. If this were not the case, the clusters would only include neighbourhoods in one city. However, most clusters included a relatively even balance of neighbourhoods from both cities.

The one exception is the cluster (the Social City Centre) associated with bars, cafes, and coffee shops. Amsterdam appeared to have many more neighbourhoods like this compared to Rotterdam. But, based on my experiences with Amsterdam, this makes sense. Amsterdam is a large tourist destination, so more neighbourhoods cater to those more touristy lifestyles, ie drinking at bars, relaxing at cafes, and smoking at coffee shops.

## Would I enjoy Rotterdam?

Based on the overall results Rotterdam appears relatively similar to Amsterdam. Therefore, I conclude that I would likely enjoy living in Rotterdam too.

## Which neighbourhood in Rotterdam is best?

From personal experience in the Amsterdam neighbourhoods, I can determine which similar areas in Rotterdam would be the most suitable for me to live in. For my year in Amsterdam I lived on the outskirts of the city, where there was plenty of outdoor space and nature, as well as easy access to many shops. Also, very conveniently there was a train station nearby.

By taking the types of clusters around where I lived in Amsterdam and comparing them to different parts of Rotterdam, areas in Rotterdam with a similar composition of cluster types can be found. That way a group of neighbourhoods similar to the ones I lived in can be found.

According to the results for the top 5 most common venue type, the best place for me to live in Rotterdam would be Kreekhuizen. A closer look at Kreekhuizen on a map reveals a beautiful large park a block away from the centre of the neighbourhood. There are also many sports fields nearby and even a large train station.

According to the results for the top 3 most common venue type, the best place for me to live in Rotterdam would be Oud-IJsselmonde. Surprisingly, this neighbourhood is quite close to Kreekhuizen, 2 kilometres away by bicycle. This neighbourhood is on the banks of the river Nieuwe Maas; there is a nature preserve island in this neighbourhood too.

# Conclusion

Based on my own personal circumstances I defined a set of questions to answer using machine learning tools taught during the IBM Data Science Professional Certificate Program. These questions allowed me to apply k-means clustering to a new set of problems.

K-means clustering was performed on two versions of the venue data for neighbourhoods in Amsterdam and Rotterdam. The two sets of data were based on the top 5 and top 3 most common types of venues for each neighbourhood.

Based on the results of the k-means clustering process, Amsterdam and Rotterdam appeared fairly similar in nature, though more neighbourhoods in Amsterdam had more of a tourism focus, as expected.

Personally, I was able to find a neighbourhood in Rotterdam that appears remarkably similar to the one I lived in during my time in Amsterdam.

Other individuals seeking to move to new cities can follow the process used in this project to gain a better understanding of the city they will be moving to, and find a new neighbourhood that they would enjoy living in. Companies in Rotterdam can also use the results of this study to help persuade international talent living in Amsterdam to move to Rotterdam, given how the two cities are largely quite similar.

Overall, I would be happy to live in either neighbourhood recommended by this process and enjoyed working on this project.

## Limitations

The process of working on a longer form project did reveal some real-world issues with data gathering. While Foursquare provided sufficient venue data for the exercises completed during the Data Science program, it did not provide enough data for all the neighbourhoods examined in this project. There was sufficient data to continue with the analysis after a few adjustments, moving to using the top 5 and 3 most common types of venues rather than top 10.

However, given the nature of the Foursquare data, individuals seeking to replicate this project for their own cities might find a severe lack of data for their city. This may occur if their city is rather small or has few Foursquare users who are creating data for venues around the city. Such individuals may need to search other venue databases for more data about their city.

An additional limitation found is in the results of this project. Based on three different methods to find the optimal number of clusters, no consensus could be found. Further, the Silhouette Scores for all numbers of clusters tried were not very high. This would indicate the clustering was not very well grouped. Given the nature of real-world data, it is simply very likely that very distinct and clear clusters could not be found. For example, if many neighbourhoods in Amsterdam and Rotterdam are very similar in terms of the types of venues in them, then the clustering algorithm will struggle to differentiate between neighbourhoods. Not much can be done about this issue, it is simply the nature of how different neighbourhoods are in the real world. For example, in Cape Town, South Africa, neighbourhoods tend to be clearer in what venue types are in them. Many housing neighbourhoods will simply not have many stores or much else other than people’s homes. Similarly, in industrial zones, there will be very few front-facing stores.

## Directions for future research

In terms of further research, a more in-depth study could be done of the neighbourhoods between cities, covering more than just the most common types of venues in each neighbourhood.

This could be in many different directions, accounting for the quality of service at different venues, or other characteristics of those venues. For example, an individual might prefer larger parks to smaller ones.

Additionally, when determining where to live exactly, more research could be performed to develop a system that accounts for additional personal preferences or requirements. For example, an individual might be happy to travel 3 kilometres to the nearest form of public transport from home, and not much further. Or they might really love going to the cinema, and only consider the neighbourhoods that have one.

# References

Coordinates Finder. (2020, 11 12). *Coordinates Finder*. Retrieved from Coordinates Finder: www.coordinatesfinder.com

Datanovia. (2020, November 11). *Determining the optimal number of clusters: 3 must know methods*. Retrieved from Datanovia: https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/

Foursquare. (2020, 11 12). *Foursquare*. Retrieved from Foursquare: https://foursquare.com/

Geeks for Geeks. (2019, 6 6). *Elbow Method for optimal value of k in KMeans*. Retrieved from geeksforgeeks: https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/

Geonames. (2020, 11 5). *Netherlands - 10 Largest Cities*. Retrieved from Geonames: https://www.geonames.org/NL/largest-cities-in-netherlands.html

Holland.com. (2020, 11 5). *Rotterdam*. Retrieved from Holland.com: https://www.holland.com/global/tourism/destinations/rotterdam.htm

Mahendru, K. (2019, 6 17). *How to Determine the Optimal K for K-Means?* Retrieved from Medium: https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb

Wikipedia. (2020, 11 5). *Category:Neighbourhoods of Rotterdam*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Category:Neighbourhoods\_of\_Rotterdam

Wikipedia. (2020, 11 12). *Neighbourhoods of Amsterdam*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Category:Neighbourhoods\_of\_Amsterdam

Wikipedia. (2020, November 12). *Wikipedia*. Retrieved from Wikipedia: https://nl.wikipedia.org/wiki/Nooddorp