Capstone Project – Battle of the neighbourhoods

Compare cities in the Netherlands.

Introduction: Business Problem

The Netherlands is a densely populated country. This results in a lack of affordable houses in different cities. As space is scarce the prices for houses is increasing in those areas. People often live in houses that are smaller than they are comfortable with because they like the city for the restaurants and bars. The current pandemic however has shown people that they can work from home and do not need to live close by their offices. This opens opportunities as the Netherlands is also a small country where distances are a matter of hours driving maximum.

Would it not be great to compare the city that you live in with other cities in the Netherlands based on same comfort in restaurants and bars but with a lower housing price?

This project aims to cluster the cities in the Netherlands based on restaurants and bars and then rank them within their cluster based on the square meter price for houses. This will then provide the needed information. You can search on your city within the list and then check which group your city belongs to. Then you will be able to see in the group what cities are cheaper than your own city.

Data

The data that is needed to answer the question on comparable cities with lower house prices:

- A list of biggest cities in the Netherlands
- The latitude and Longitude of those cities
- The square meter price for houses
- The categories of interest on which to compare the cities

A list of the biggest cities in the Netherlands can be retrieved from https://wikikids.nl/Lijst_van_grote_Nederlandse_steden. This site holds a list with all cities with more than 50.000 residents plus one other city (Middelburg) as that is the capital of a county (Zeeland).

The latitude and longitude can be added to this list by using openstreetmap data which is accessed via GeoPy geocoder Nominatim.

The square meter sales prices for houses can be found on HuizenZoeker.nl. An excel file with information on prices including the house prices per square meters over the last years can be retrieved from https://www.huizenzoeker.nl/woningmarkt/download/woningmarkt-

gemeenten.xls?d=20210221. It holds the prices for each month. In this project the data from January 2021 will be used.

The categories on which the data will be grouped can be found on https://developer.foursquare.com/docs/build-with-foursquare/categories/ . In this project the focus will be on restaurants and pubs. The restaurants are specified in more detail to search for Cajun, Creole, Indonesian and Mediterranean cuisine.

The cities will be compared based on the top 15 categories that are returned by foursquare.

Data Cleaning

The data in the different list can be combined but some of it is not used in this project. There are four cities listed in list of biggest cities that does not match the list of housing prices. This due to different ways of entering this data. The four cities that will be changed to match the data are:

- "'s-Gravenhage" which is also known as "Den Haag" => "Den Haag" will be used as city name.
- "'s-Hertogenbosch" which is also known as "Den Bosch" => "Den Bosch" will be used as city name.
- "Spijkenisse" is named in the list of cities but the municipal "Nissewaard" is mentioned in the list with prices. => "Spijkenisse" will be used as city name.
- "Hengelo" is listed as "Hengelo (O)"in the list with house prices. => "Hengelo" will be used as city name.
- "Westland" is the municipal but the biggest city in that municipal has only 17.000 residents so this line will be dropped.

The list of house prices contains data in different sheets and with different columns. In this project the average house price per square meter (Gem. vraagprijs per m2) will be used from the month January 2021.

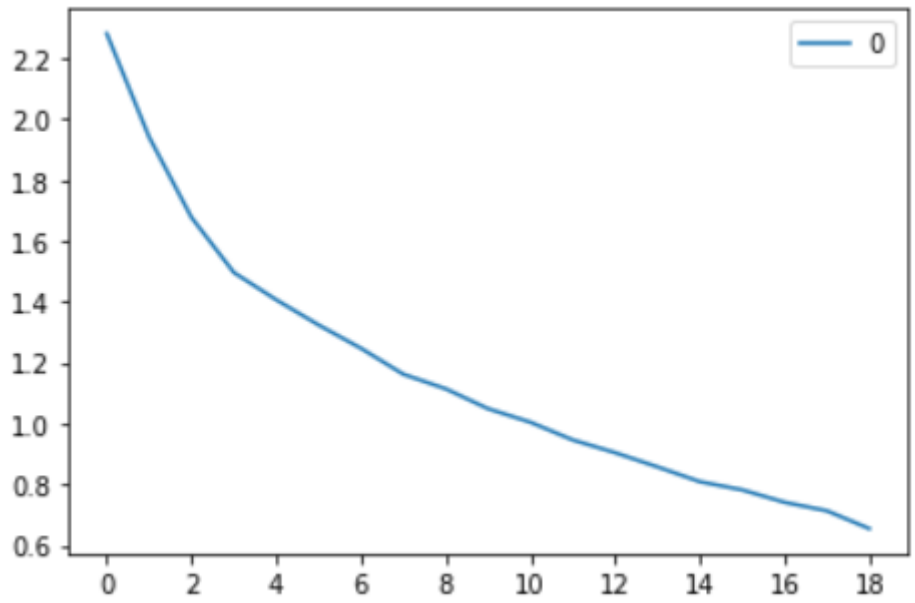
Methodology

In this project we will focus on detecting similarity between cities based on restaurants and bars. We will limit our analysis to an area 5km around the city centre.

In first step we have collected the required data: location, average house prices per square meters and category of the restaurants and bars within 5km from the city centre (according to Foursquare categorization). We have created an overview of the top 10 categories for the biggest cities in the Netherlands.

Second step in our analysis will be the use of unsupervised clustering by means of the **k-means clustering**. To be able to use this method we will need to identify the optimal number of clusters. This will be done with the elbow method.

First step is to define a function to run the K-means with different parameters. Then the outcome of these runs will be plotted to show the result. Based on the plotted data we should be able to use the elbow method to define the optimal number of clusters.

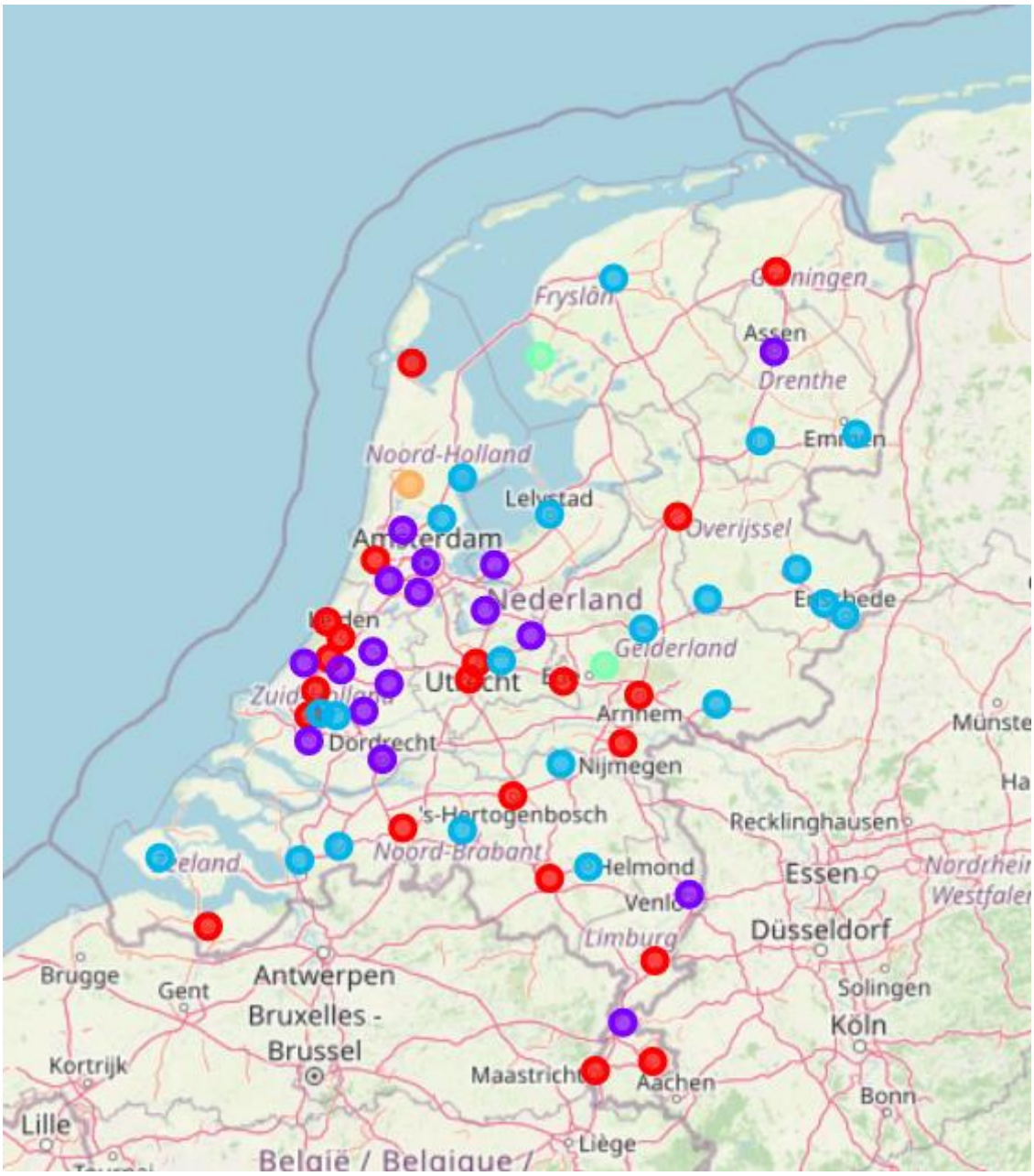


As the elbow method is not immediately clear from the plot, we calculated the number of clusters which resulted in 5.

Result

The result of the clustering is a dataframe sorted by cluster. Within that cluster it is sorted by Average house price per square meter from most expensive to least expensive city.

After that, the results are shown on a map of the Netherlands with different colours. Selecting one of the cities on the map will show the city, the cluster and the average price per square meter in January 2021 for a house in that city.



Discussion

The project concentrated on checking if it was feasible to compare cities based on preferences in restaurants and pubs. In the project multiple factors were changed like the number of categories and the number of top categories per city to see what impact they had. In the end 10 categories were chosen to query the Foursquare API. Next to that the top 10 categories were listed to cluster upon. Increasing this number of categories led to more clusters with only one or two cities within the cluster.

This idea can be enhanced for future purposes. If an end user could select the top 10 categories that (s)he finds the most important this would generate a result based on that input.

Conclusion

The conclusion of this project is that it is possible to find cities that are comparable to your own city. Depending on which city you live in you can find cheaper and more expensive cities to which you can

move. In my case I live in Delft which belongs to cluster 0. This means that the cheapest comparable city would be Heerlen and the most expensive comparable city is Haarlem.