Living in the City of Eindhoven

Capstone project by Peter van Liesdonk

Introduction

For the last few years I have been living in the city of Eindhoven, a small provincial town in the south of the Netherland and mostly known for its high tech industries. For this reason it is sometimes called Silicon Valley of the Netherlands. I like living here, and I'm living in a great neighbourhood. Unfortunately I cannot stay here: I'm currently residing in a social housing project, and I'm getting too rich to stay here. I know: luxury problem.

So where in this city will I buy a new house?

To figure this out I want to compare the various neighbourhoods in Eindhoven to see how similar they are, and finally figure out if there is a neighbourhood similar to mine where I could also live. I will do this based on the different venues and ammenities available in the direct vicinity of the various neighbourhoods.

As in the course, I'd like to cluster neighbourhoods on similarity and show them on a map. Then I'd also like to create a decision tree that can show me the most important factors for choosing a certain neighbourhood.

Data

To know more about Eindhoven I need as much data as possible. I found the following interesting datasets:

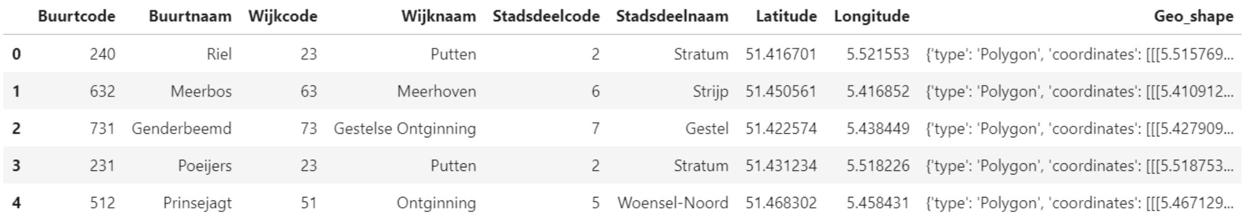
- A list of neighbourhoods in Eindhoven at [https://data.eindhoven.nl/explore/dataset/buurten/export/]. This includes
  - Name of neighbourhood (buurt), residential areas (wijken) and boroughs (stadsdeel) within Eindhoven,
  - their geographic coordinates,
  - their borders in GeoJSON format.
- A table of key figures about the various neighbourhoods [https://opendata.cbs.nl/statline/#/CBS/nl/dataset/84286NED/table?ts=1546775064672], which includes things like
  - population,
  - population density,
  - area,
  - amount of house

In addition I will use Foursquare to find popular venues close to each of the neighbourhoods. Using Foursquare might mean that I do not actually need the last of the tables above. Another difficulty might be the sparsity of information on Foursquare, since Foursquare is not very popular in the Netherlands.

Methodology

Exploring the dataset

We start by importing the first dataset, a list of neighbourhoods and their geographic position from [https://data.eindhoven.nl/explore/dataset/buurten/export/]. After importing in Pandas and cleaning up some of the column names we get a dataset that looks as follows:



	Buurtcode	Buurtnaam	Wijkcode	Wijknaam	Stadsdeelcode	Stadsdeelnaam	Latitude	Longitude	Geo_shape
0	240	Riel	23	Putten	2	Stratum	51.416701	5.521553	{'type': 'Polygon', 'coordinates': [[[5.515769...
1	632	Meerbos	63	Meerhoven	6	Strijp	51.450561	5.416852	{'type': 'Polygon', 'coordinates': [[[5.410912...
2	731	Genderbeemd	73	Gestelse Ontginning	7	Gestel	51.422574	5.438449	{'type': 'Polygon', 'coordinates': [[[5.427909...
3	231	Poeijers	23	Putten	2	Stratum	51.431234	5.518226	{'type': 'Polygon', 'coordinates': [[[5.518753...
4	512	Prinsejagt	51	Ontginning	5	Woensel-Noord	51.468302	5.458431	{'type': 'Polygon', 'coordinates': [[[5.467129...

Imporing the key figures dataset from [https://opendata.cbs.nl/statline/#/CBS/nl/dataset/84286NED/table?ts=1546775064672] gives us the following columns.

```
['Wijken en buurten', 'Gemeentenaam', 'Soort regio', 'Codering',
'Indelingswijziging wijken en buurten', 'Inwoners', 'Inwoners 15 tot 25
jaar', 'Inwoners Westers totaal', 'Inwoners Nederlandse Antillen en Aruba',
'Eenpersoonshuishoudens', 'Bevolkingsdichtheid', 'Woningvoorraad',
'Percentage meergezinswoning', 'Personenauto's; brandstof benzine',
'Motorfietsen', 'Oppervlakte', 'Mate van stedelijkheid',
'Omgevingsadressendichtheid']
```

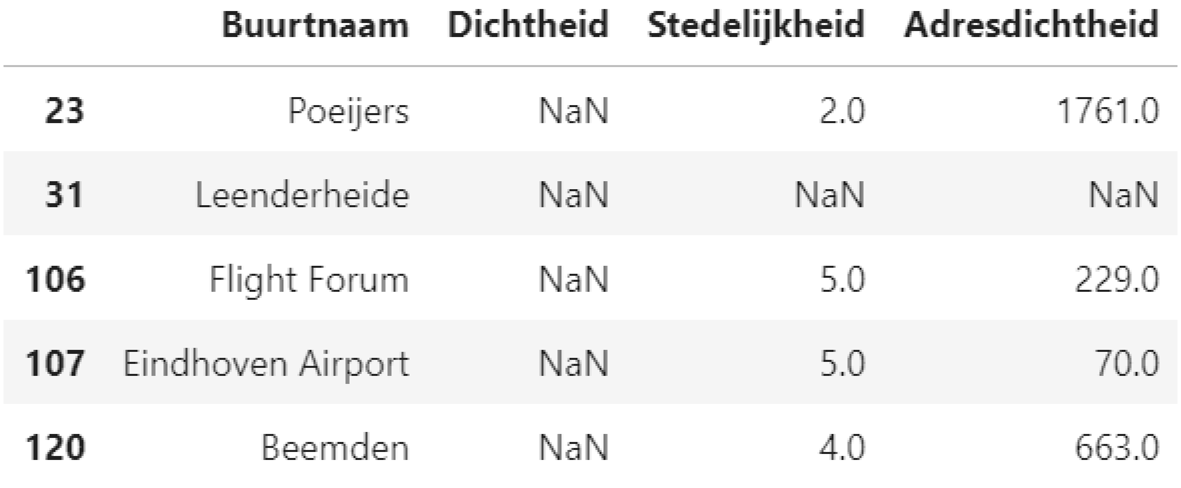
This requires some more cleanup.

First of all we only want neighbourhoods (wijken) while this dataset also has aggregated information. This can be done by filtering on the 'Soort Regio' column.

Secondly, not all data is relevant. We decide to filter only the following columns, since they seem to be the most relevant to the problem at hand:

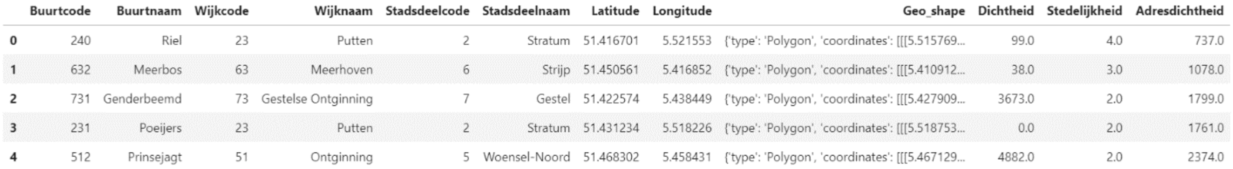
- Wijken en buurten: the neighbourhood name
- Bevolkingsdichtheid: population density.
- Mate van stedelijkheid: Urbanization, how city-like is the neighbourhood
- Omgevingsadressendichtheid: How many addresses per square km

After cleanup we get the following results that still gave NaN values:



	Buurtnaam	Dichtheid	Stedelijkheid	Adresdichtheid
23	Poeijers	NaN	2.0	1761.0
31	Leenderheide	NaN	NaN	NaN
106	Flight Forum	NaN	5.0	229.0
107	Eindhoven Airport	NaN	5.0	70.0
120	Beemden	NaN	4.0	663.0

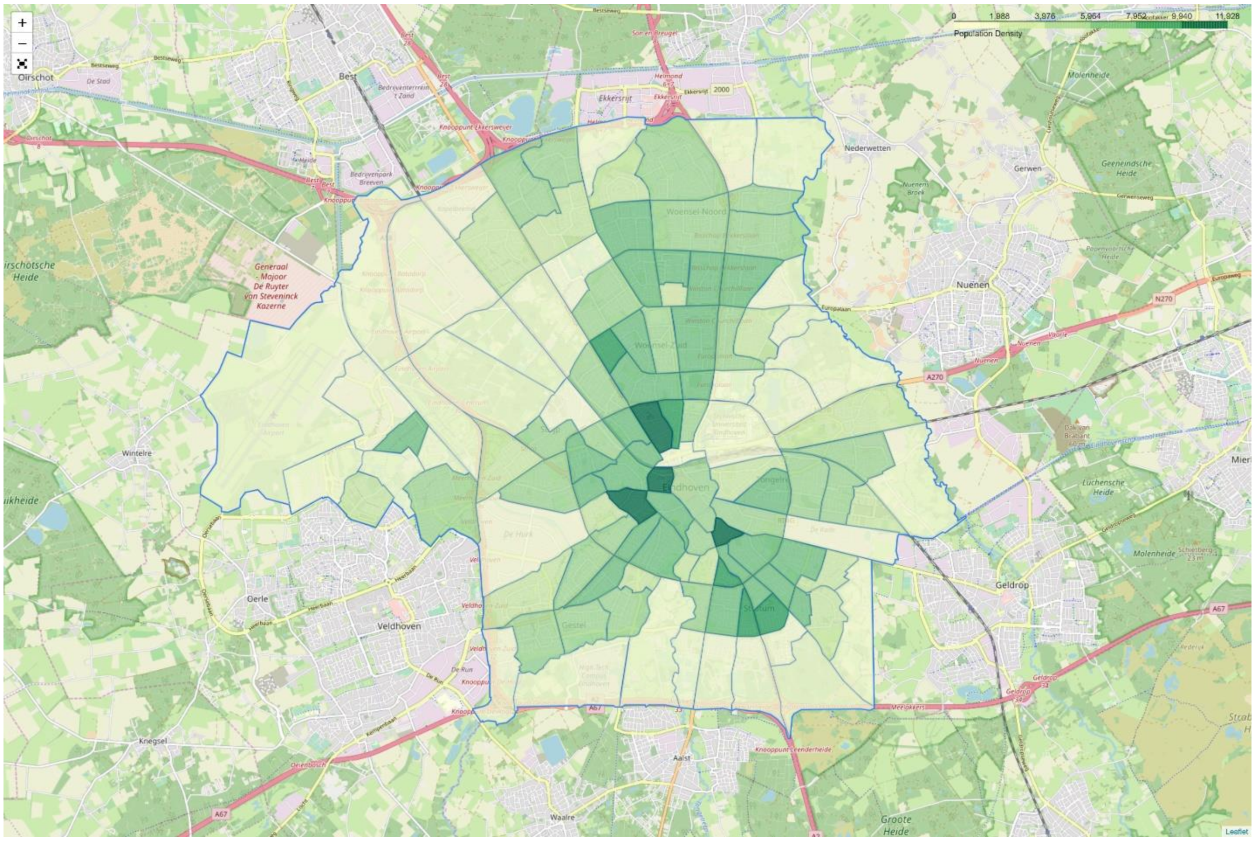
We know all of these are industrial areas, and replace the NaNs with 0. When we merge with the previous dataset, we get the following result:



	Buurtcode	Buurtnaam	Wijkcode	Wijknaam	Stadsdeelcode	Stadsdeelnaam	Latitude	Longitude	Geo_shape	Dichtheid	Stedelijkheid	Adresdichtheid
0	240	Riel	23	Putten	2	Stratum	51.416701	5.521553	{'type': 'Polygon', 'coordinates': [[[5.515769...	99.0	4.0	737.0
1	632	Meerbos	63	Meerhoven	6	Strijp	51.450561	5.416852	{'type': 'Polygon', 'coordinates': [[[5.410912...	38.0	3.0	1078.0
2	731	Genderbeemd	73	Gestelse Ontginning	7	Gestel	51.422574	5.438449	{'type': 'Polygon', 'coordinates': [[[5.427909...	3673.0	2.0	1799.0
3	231	Poeijers	23	Putten	2	Stratum	51.431234	5.518226	{'type': 'Polygon', 'coordinates': [[[5.518753...	0.0	2.0	1761.0
4	512	Prinsejagt	51	Ontginning	5	Woensel-Noord	51.468302	5.458431	{'type': 'Polygon', 'coordinates': [[[5.467129...	4882.0	2.0	2374.0

Putting Eindhoven on a map

We can now compile a simple map that shows the different neighbourhoods in a Choropeth map, showing the population density.



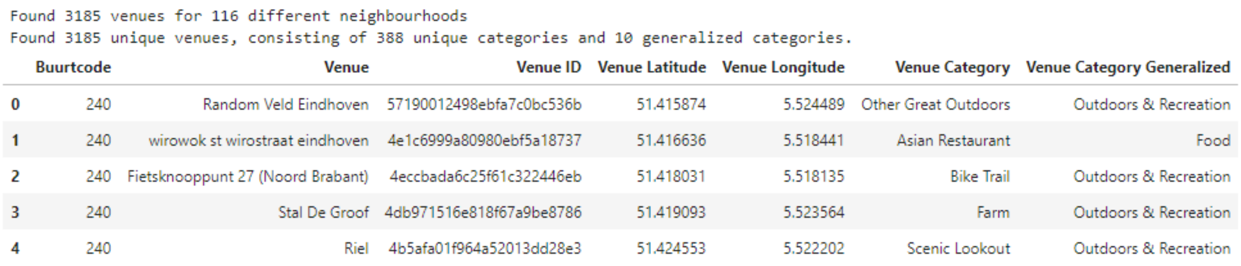
Venues from Foursquare

Next, we want to get a list of interesting venues from Foursquare. We want to figure out the venues in each neighbourhood. This is difficult, as Foursquare allows us to search around around specific coordinates with no option of restricting to a neighbourhood.

We solve this as follows: using the coordinates and the geometric shape we can find the smallest circle that completely surrounds the neighborhood. We will use the radius of this circle to restrict the Foursquare searches.

In addition, Foursquare gives us too many different venue categories. We generalize every category to one of 10 main categories.

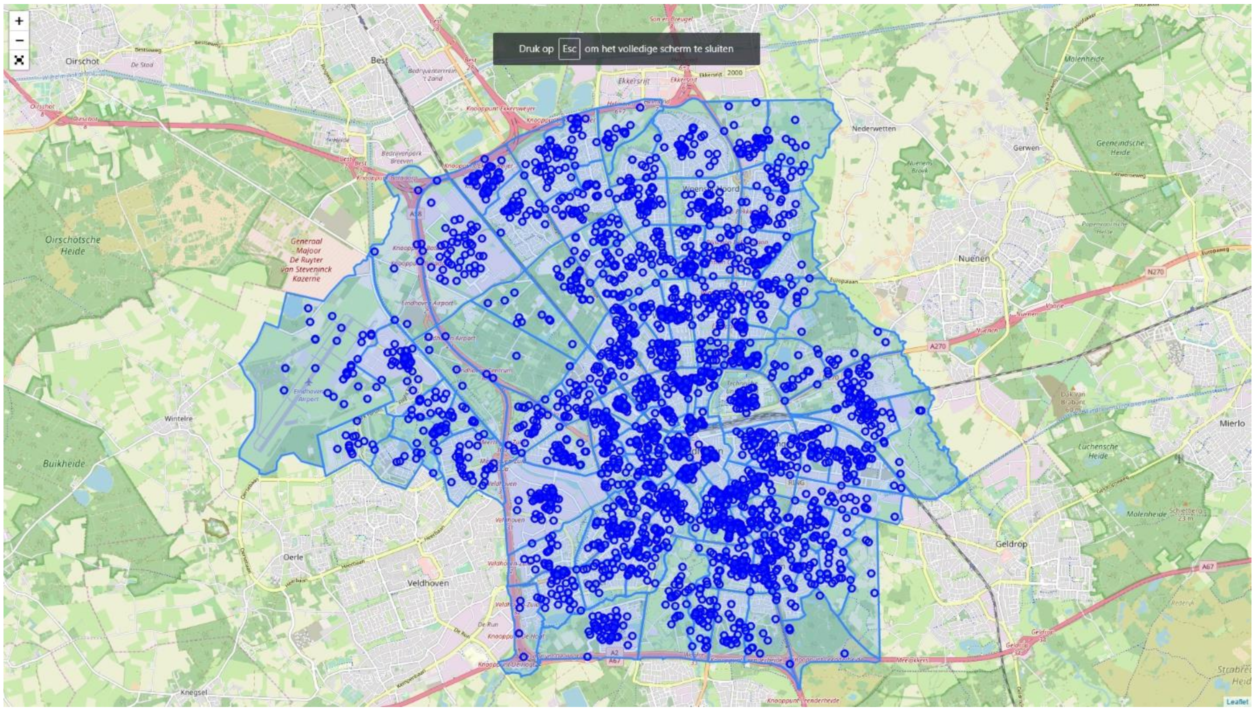
This gives us a dataframe of venues as follows. This frame contains many duplicates, which are venues that fall within the readius of multiple neighbourhoods.



```
Found 3185 venues for 116 different neighbourhoods
Found 3185 unique venues, consisting of 388 unique categories and 10 generalized categories.
```

	Buurtcode	Venue	Venue ID	Venue Latitude	Venue Longitude	Venue Category	Venue Category Generalized
0	240	Random Veld Eindhoven	57190012498ebfa7c0bc536b	51.415874	5.524489	Other Great Outdoors	Outdoors & Recreation
1	240	wirowok st wirostraat eindhoven	4e1c6999a80980ebf5a18737	51.416636	5.518441	Asian Restaurant	Food
2	240	Fietsknooppunt 27 (Noord Brabant)	4eccbada6c25f61c322446eb	51.418031	5.518135	Bike Trail	Outdoors & Recreation
3	240	Stal De Groof	4db971516e818f67a9be8786	51.419093	5.523564	Farm	Outdoors & Recreation
4	240	Riel	4b5afa01f964a52013dd28e3	51.424553	5.522202	Scenic Lookout	Outdoors & Recreation

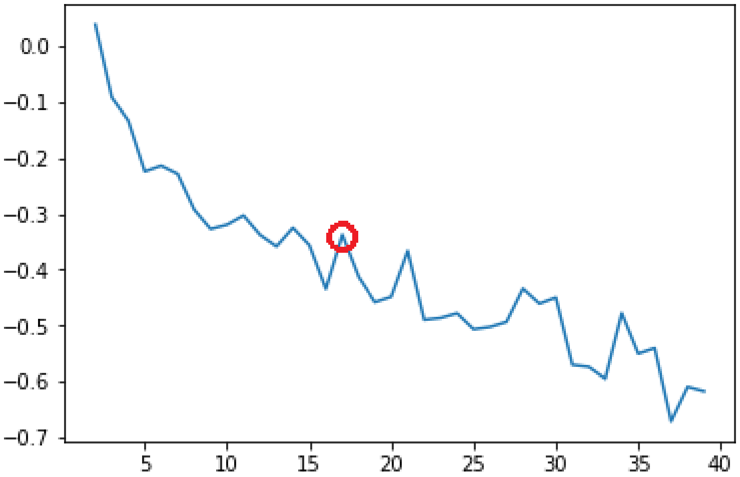
We can put all the found venues on the map as follows:



Clustering Neighborhood

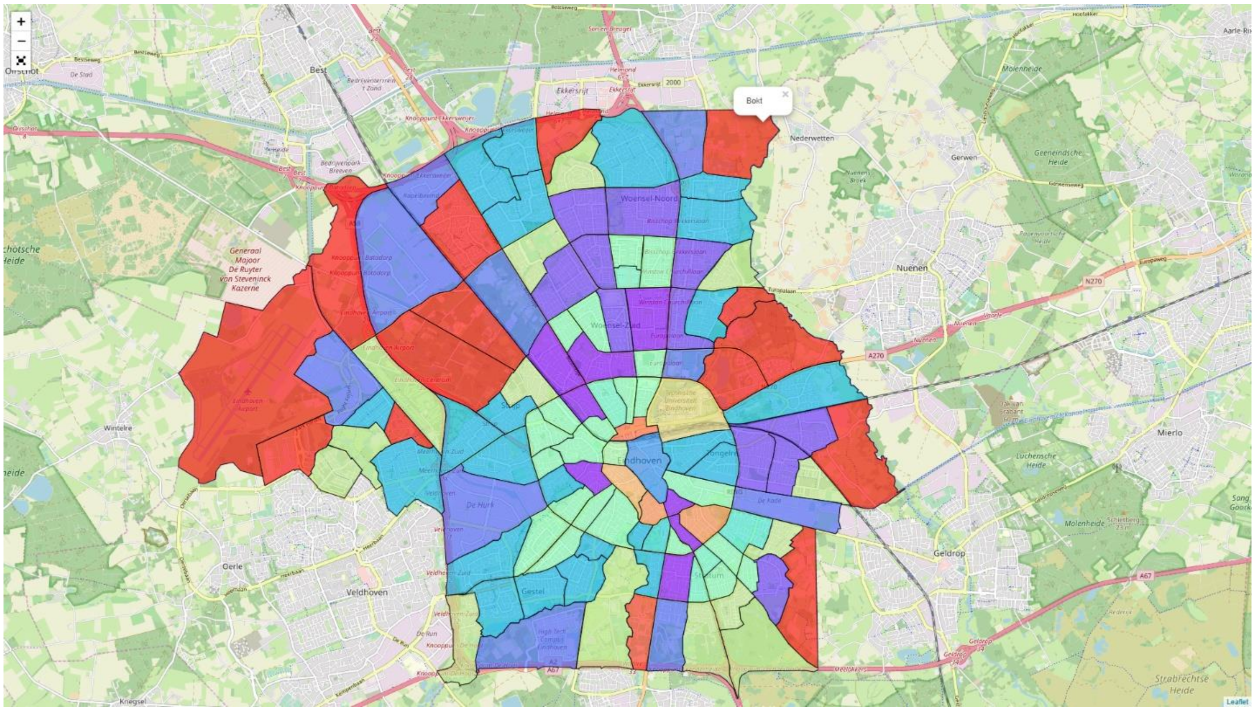
Next, we want to cluster all of the neighborhoods. We will be using a K-means clustering algorithm. In order to do this, we first turn the venues dataframe into one-hot notation and sum how often each venue occurs for each neighborhood. To this we add population density, urbanization and housing density figures.

To find an optimal value for k, we run the clustering algorithm for various values and plot the silhouette score.



We immediately see there is no clear choice for . Choosing (the maximum) doesn't give any real insight. We decide to go for , which is a local optimum and might provide interesting insights.

This results in the following clustering:



Decision Tree

Finally we want to build a decision tree in order to figure in which type of neighbourhood I should live.

We train a decision tree using the existence of certain venues as training data in order to predict the cluster labels we found above. This gives the following result:

Results

When we analyze the results of the clustering, we get the following results:

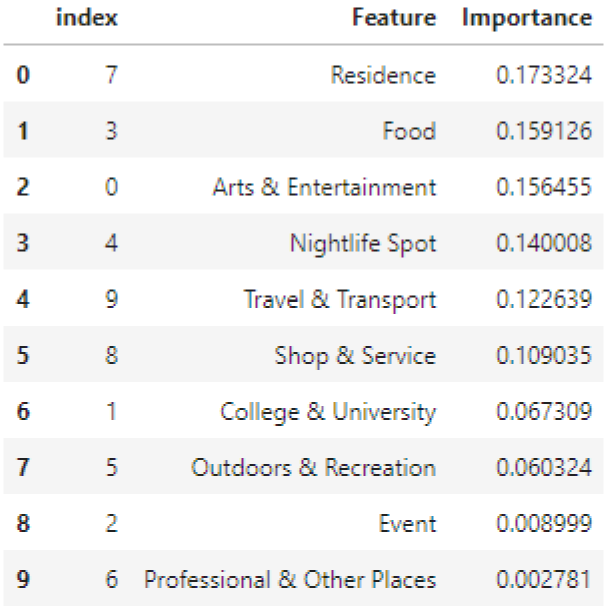
Cluster	Count	Neighbourhoods
0	14	Riel, BeA2, Urkhoven, Bokt, Bosrijk, Karpen, Wielewaal, Herdgang, Park Forum, Eindhoven Airport, Koudenhoven, Kerkdorp Acht, Eckartdal, Castiliëlaan
1	8	Woensel-West, Eliasterrein, Vonderkwartier, Joriskwartier, Kronehoef, Gerardusplein, Generalenbuurt, Limbeek-Noord, Rochusbuurt
2	11	Prinsejagt, Jagershoef, Tempel, Doornakkers-West, Oude Gracht-West, Gijzenrooi, Woenselse Heide, Eckart, Doornakkers-Oost, Rapenland, Muschberg, Geestenberg

Cluster	Count	Neighbourhoods
3	10	Poeijers, Mispelhoef, Hurk, Kapelbeemd, Esp, Beemden, Flight Forum, Hondsheuvels, Vredeoord, Eikenburg
4	1	Binnenstad
5	23	Genderbeemd, Lievendaal, 't Hofke, Hanevoet, Bennekel-West, Gagelbosch, Achtse Barrier-Hoeven, Rapelenburg, Elzent-Zuid, Grasrijk, Achtse Barrier-Spaaihoef, Achtse Barrier-Gunterslaer, Het Ven, Heesterakker, Drents Dorp, Oude Gracht-Oost, Bennekel-Oost, Vaartbroek, Blixembosch-Oost, Lakerlopen, Zandrijk, Engelsbergen, Genneperzijde, Villapark
6	1	Looiakkers
7	1	Burghplan

Cluster	Count	Neighbourhoods
8	4	Winkelcentrum, Nieuwe Erven, Hemelrijken, Kerstroosplein
9	22	Schoot, Strijp S, Kruidenbuurt, Sintenbuurt, Irisbuurt, Schouwbroek, Mensfort, Philipsdorp, Blaarthem, 't Hool, Hagenkamp, Oude Toren, Barrier, Vlokhoven, Limbeek-Zuid, Tivoli, Bloemenplein, Oude Spoorbaan, Woenselse Watermolen, Genderdal, Schrijversbuurt, Gildebuurt
10	1	Witte Dame
11	14	Meerbos, Gennep, Luytelaer, Waterrijk, Meerrijk, Puttense Dreef, Zwaanstraat, Ooievaarsnest, Leenderheide, Roosten, Blixembosch-West, Tongelresche Akkers, Schuttersbosch, Driehoeksbos

Cluster	Count	Neighbourhoods
12	1	TU-terrein
13	1	Bergen
14	2	Tuindorp, Elzent-Noord
15	1	Fellenoord
16	1	Sportpark Aalsterweg

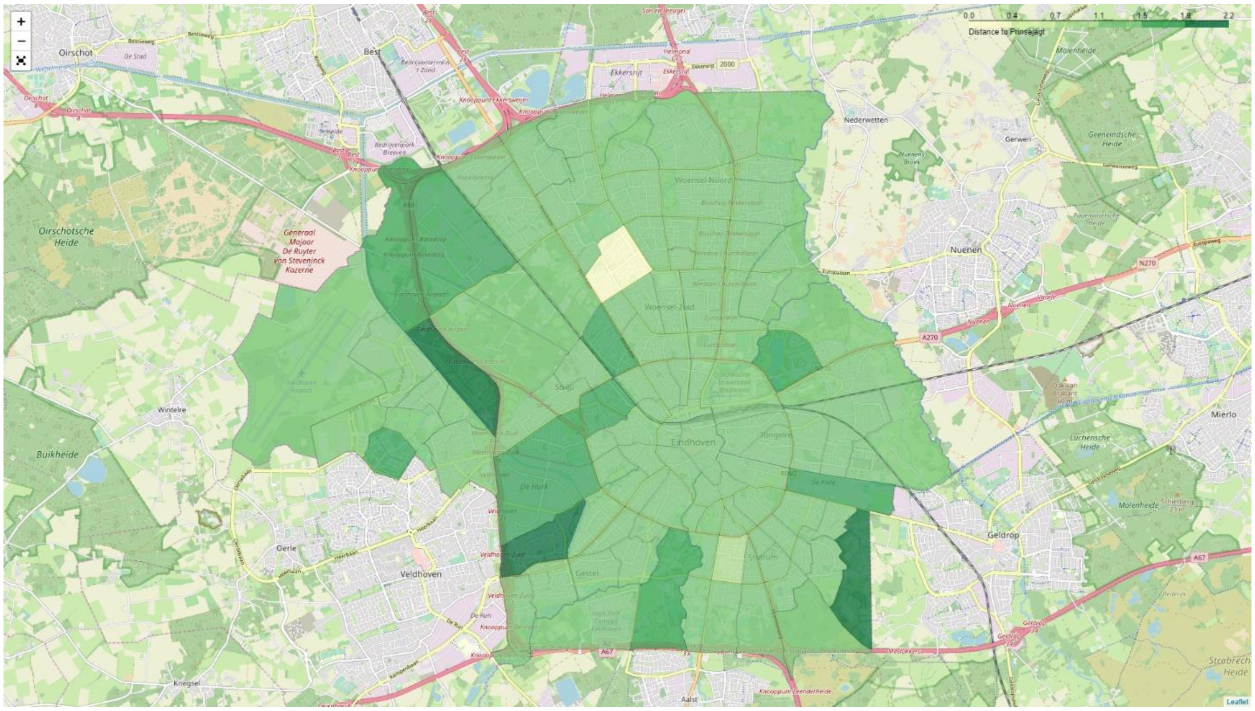
It is also interesting to see what the most distinguishing features are between neighbourhoods. We can distill this information from the decision tree we built:



	index	Feature	Importance
0	7	Residence	0.173324
1	3	Food	0.159126
2	0	Arts & Entertainment	0.156455
3	4	Nightlife Spot	0.140008
4	9	Travel & Transport	0.122639
5	8	Shop & Service	0.109035
6	1	College & University	0.067309
7	5	Outdoors & Recreation	0.060324
8	2	Event	0.008999
9	6	Professional & Other Places	0.002781

From this we can deduce that the biggest difference between clusters is the presence of entertainment and nightlife. It appears

We also wanted to look how different each neighbourhood is from my current neighbourhood Prinsejagt. We used the same data as input into the clustering algorithm, but now just look at the difference from Prinsejagt.

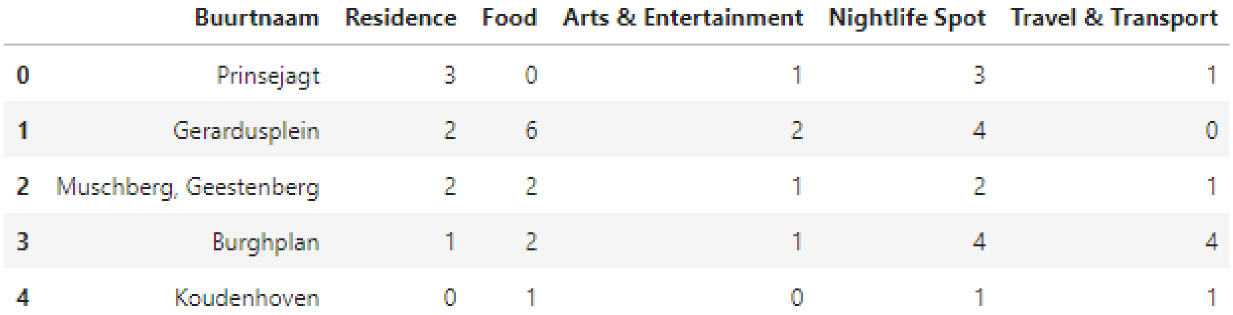


We can sort the neighbourhoods by their distance:

Rank	Neighbourhood	Distance
1	Prinsejagt	0.00
2	Gerardusplein	0.97
3	Muschberg, Geestenberg	1.13
4	Burghplan	1.15

Rank	Neighbourhood	Distance
5	Koudenhoven	1.17

Finally, let's correlate these neighbourhoods with the most distinguishing features.



	Buurtnaam	Residence	Food	Arts & Entertainment	Nightlife Spot	Travel & Transport
0	Prinsejagt	3	0	1	3	1
1	Gerardusplein	2	6	2	4	0
2	Muschberg, Geestenberg	2	2	1	2	1
3	Burghplan	1	2	1	4	4
4	Koudenhoven	0	1	0	1	1

Discussion

Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.

There are some obvious clusters in here, based on my subjective experience:

- Cluster 0 contains lots of industrial neighbourhoods with little people living there
- The clusters with only one neighbourhood all indicate a quite special neighbourhood, easily distinguished.
- Clusters 1, 2, 3, 5 and 7 consists of neighbourhoods with lot of houses
- Cluster 3 contains shops
- Cluster 8 is a university campus
- Cluster 9 has a almost no buildings

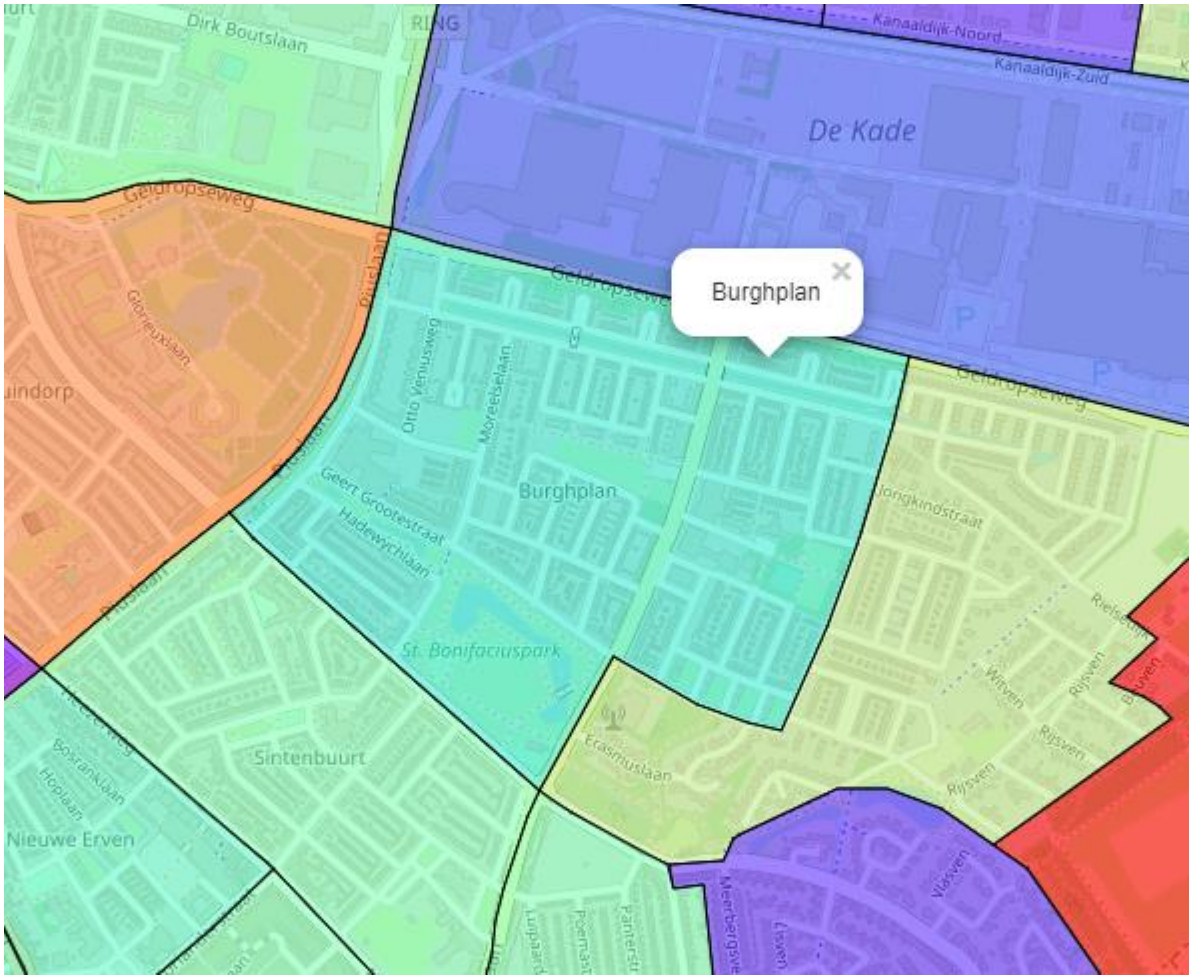
Second obvious thing we notice is how the ranking of distance to Prinsejagt differs from the clustering, even though they cover four different clusters. This is due to the different distance mechanisms: K-means optimizes in-cluster inertia, which apparently gives different results from regular euclidean distance.
Finally we took a closer look at the amount of distinguishing features for each of the neighbourhoods closest to Prinsejagt. This seems to indicate that there is still quite a difference between these neighbourhoods
.

Conclusion

The analysis above indicates that the neighbourhoods of Gerardusplein, Muschberg/Geestenberg, Burghplan or Koudenhoven are alternatives to Prinsejagt. Each of these is worth visiting an checking out.

The data also seems to indicate that Burghplan is better connected, with better nightlife spots. This makes Burghplan my first option to look for new housing.

```
Buurtcode                                                       23
2
Buurtnaam                                                 Burghpla
n
Wijkcode                                                         2
3
Wijknaam                                                     Putte
n
Stadsdeelcode
2
Stadsdeelnaam                                               Stratu
m
Latitude                                                    51.427
8
Longitude                                                   5.5069
3
Geo_shape         {'type': 'Polygon', 'coordinates': [[[5.513998..
.
Dichtheid                                                      610
0
Stedelijkheid
2
Adresdichtheid                                                 218
1
Common Venue 1                          Professional & Other Place
s
Common Venue 2                                       Shop & Servic
e
Common Venue 3                                Outdoors & Recreatio
n
Common Venue 4                                   Travel & Transpor
t
Common Venue 5                                       Nightlife Spo
t
Cluster label
7
d_prinsejagt                                                1.1469
6
Name: 53, dtype: object
```


Burghplan
De Kade
Dirk Boutslaan
RING
Kanaaldijk-Noord
Kanaaldijk-Zuid
Geldropseweg
Geldropseweg
Geldropseweg
Gloriëuxlaan
uindorp
Otto Veniusweg
Moreeselaan
Burghplan
Geert Grootestraat
Hadewychlaan
St. Bonifaciuspark
Jongkindstraat
Rielsedijk
Witven
Rijsven
Rijsven
Rijsven
Erasmuslaan
Sintenbuurt
Bosranklaan
Hoplaan
Nieuwe Erven
Vlasven
Lisven
Meerbergsven