Introduction

Eindhoven is one of the fastest growing tech hubs in Europe. It has been dubbed “the Brainport” and has attracted talent from all over the world. This growing population will bring with it the possibility of higher economic output. Companies then would want to expand locally into new areas to capture more of that business. Eindhoven however, has different neighbourhoods with distinct cultures and amenities. To find the right neighbourhoods or areas to expand into, we therefore need to identify which areas are the most similar.

To find the most similar areas, this project will use K-Means clustering to find the neighbourhoods with the most similar venues based on their category and frequency.

Data

To find the various areas of the city and its venues, two data sources are used. The first source is the local directory which has data about the various areas. From this source, we will use the Dutch postal code number to identify different areas. The Dutch postal code system breaks down postal code areas in codes of four numbers and two letters. The four numbers define the area within a municipality, while the letters break it down even further. For the purposes of this project, we will only focus on the four letter code.

The second data source, is the Foursquare API which will return all venues together with their corresponding postal code. Combining the Foursquare data with the area data, we can find locations where the overall venue offerings are similar.

From the area data, we drop the columns containing redundant information. These include the streetname, alternative postal code, neighbourhood code and other variables. In total, there are 34 unique postal codes in the list. Below, an overview of the extracted dataframe:

	Postcode	Neighborhood	District Code	District Name	Latitude	Longitude
0	5612.0	Limbeek-Noord	4	Stadsdeel Woensel-Zuid	51.450079	5.462462
1	5641.0	't Hofke	3	Stadsdeel Tongelre	51.448921	5.518855
2	5644.0	Kerstroosplein	2	Stadsdeel Stratum	51.419342	5.495323
3	5622.0	Barrier	4	Stadsdeel Woensel-Zuid	51.463883	5.460216
5	5624.0	Prinsejagt	5	Stadsdeel Woensel-Noord	51.463879	5.458593
6	5621.0	Woensel-West	4	Stadsdeel Woensel-Zuid	51.452878	5.456117
8	5643.0	Kruidenbuurt	2	Stadsdeel Stratum	51.419250	5.500829
9	5614.0	Tuindorp	2	Stadsdeel Stratum	51.429579	5.494769
10	5632.0	Heesterakker	5	Stadsdeel Woensel-Noord	51.484700	5.498735
11	5615.0	Bloemenplein	2	Stadsdeel Stratum	51.427301	5.486916
18	5629.0	Blixembosch-West	5	Stadsdeel Woensel-Noord	51.490877	5.465886

Figure 1: Dataframe of all Eindhoven postal codes

The Foursquare data was extracted using their API explore function. This was done with a limit of 150 venues and a radius of 2000 meters which corresponds with the average neighbourhood size in Eindhoven according to the Dutch Central Statistics.

The data retrieved is as follows (with some columns dropped for better oversight):

	name	categories	lat	lng	postalCode	city
0	Yoghurt Barn Eindhoven	Frozen Yogurt Shop	51.440187	5.478956	5611 DB	Eindhoven
1	De Bierbrigadier	Beer Store	51.436422	5.476816	NaN	Eindhoven
2	DENF Coffee	Coffee Shop	51.439165	5.474124	NaN	Eindhoven
3	De Vooruitgang	Restaurant	51.439171	5.478755	5611 EB	Eindhoven
4	The Student Hotel	Hotel	51.441995	5.480799	5611 AA	Eindhoven

Figure 2: Dataframe retrieved from the Foursquare API

The postal codes include NaN values and includes the letter values behind the letter postal codes. The values with NaN are dropped from the list and the letter values are removed from remaining postal codes. This gives us 82 total venues with 67 unique categories. Below, an overview of the unique values and their frequency.

categories	
Restaurant	6
French Restaurant	5
Bar	4
Coffee Shop	4
Brewery	3
...	...
Gastropub	1
Asian Restaurant	1
Grocery Store	1
Gym	1

Figure 3: Unique values and their frequencies

One hot encode is used to place these venues in the postal codes by transforming them into categorical values. These values are added together to see how often a certain category occurs in each postal code. Below an overview:

Postcode	city	Afghan Restaurant	Asian Restaurant	Bagel Shop	Bar	Bed & Breakfast	Beer Store	Bookstore	Breakfast Spot	Brewery	...
5654	1	1	1	1	1	1	1	1	1	1	...
5652	1	1	1	1	1	1	1	1	1	1	...
5644	1	1	1	1	1	1	1	1	1	1	...
5617	7	7	7	7	7	7	7	7	7	7	...
5616	8	8	8	8	8	8	8	8	8	8	...
5615	7	7	7	7	7	7	7	7	7	7	...
5614	1	1	1	1	1	1	1	1	1	1	...
5613	4	4	4	4	4	4	4	4	4	4	...

Figure 4: Frequency of venue categories at each postal code

In this final list, only 10 postal codes are available. This is unfortunate given the aim of this project. However, this shows the need for this analysis.

Methodology

The data from the previous section is used to find the regions with the most similarities based on their venues. To achieve this, firstly the top ten most common venues are identified for each postal code to create a profile for each. Secondly, K-means clustering is used to identify which regions are the most similar. K-means clustering uses an unsupervised approach to find regions with similar traits and segments them into different categories.

Due to the lack of venue data for all postal codes, we are limited by the 10 postal codes with Foursquare venue information. Due to this low number, the code will limit the maximum number of clusters to 5.

Results

After executing the methodology described above, the final results show that there are 5 different clusters in Eindhoven. In the image below, an overview of their location is shown.

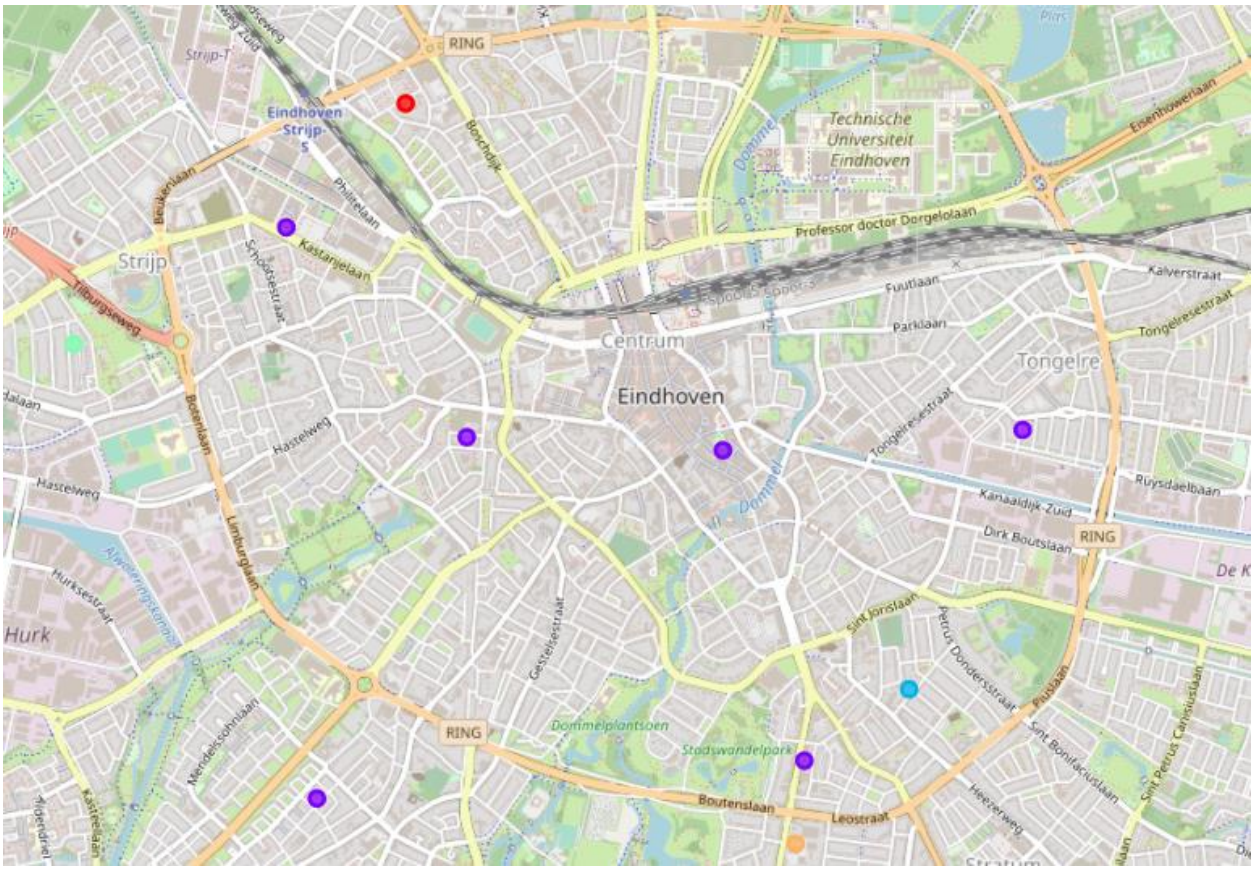

Image 1: Map of Eindhoven together with the clusters.

Discussion

From the results and the map we see that cluster 1 (dark blue dots) is clearly predominant. Regions that fall under cluster 1 are primarily focused within the city 'ring', showing that regions in the inner-city are similar to one another. The other clusters (0, 2, 3 and 4) are either outside of the city ring or are near the ring.

As for economic possibilities, these results show that within the city ring the venues are pretty similar. Expansion from one region in the ring to the next will have little difference in potential clientele. The closer your location is to the edges of the city, or beyond it, the more risk you incur due to the differing composition of the areas.

As previously mentioned, there is only data for 10 regions in Eindhoven. This is not enough to come to definitive conclusions. However, these preliminary results do show what potential conclusions can look like if we find a different source of venue data.

Conclusion

Although data is scarce and these results can change with more information, we can conclude that regions in the city ring in Eindhoven tend to be similar. The best recommendation therefore for companies located within the ring is to expand business within the ring. Expanding to areas near the edges or outside the ring will require a different strategy since they tend to be more dissimilar.