The perceived quality of neighborhood amenities

A case study set in The Hague, Netherlands
for:
IBM Applied Data Science Capstone

12 April 2020
By: Henry Veenbergen

INTRODUCTION

The Hague is a city located on the west coast of The Netherlands. With around 550,000 inhabitants, it is the country's 3rd largest city. The city houses the country's government, houses of parliament and many of the country's civil servants in the many ministries. These are mostly located in the city centre.
The heart of the city is complemented by large numbers of shops, bars and restaurants.
On the western end of the city are the Scheveningen & Kijkduin beach areas, which are very popular with tourists and locals alike
In and around all this are all the neighborhoods, which together form a highly diverse city.

PROBLEM DEFINITION

As in most cities, the neighborhoods in The Hague vary greatly, for example in terms of:

- Wealth
- Ethnic composition
- Age distribution
- Degree of urbanisaton

At the same time, each of these neighborhoods would have many of the same types of amenities, such as:

- Restaurants
- Café's & Bars
- Groceries
- Bakeries

While one would intuitively expect some of the neighborhood characteristics, most notably wealth, to affect the quality of amenities, this may not necessarily be the case. The lack of a readily available, objective way of measuring quality complicates the issue.

The question therefore is twofold:

- Does the perceived quality of amenities vary between neighborhoods?
- Can key neighborhood characteristics explain this difference?

These questions may be relevant for a range of people, including entrepreneurs looking to open a new business and (prospective) residents.

THE DATA

The data that will be used for this project comprises:

1. **Information from the Dutch National Statistics office (CBS)**
   In particular a dataset with respect to Dutch cities and neighborhoods. This dataset (named '84583NED') contains a wealth of key demographic and economic data for each neighborhood in The Hague.

This is a quite a substantial dataset with 107 columns, though not all are filled.

It can be retrieved in a number of ways, including:

- Through an API
  Once the 'cbsodata' module is installed on the PC, the module can be imported in Python and the dataset can be loaded directly into a Pandas dataframe, for example:
  Neighborhood_data = pd.DataFrame(cbsodata.get_data('84583NED'))

- Downloading the data from the CBS website.
  https://opendata.cbs.nl/statline/#/CBS/nl/dataset/84583NED/table?ts=1586113019879
  This has the advantage that a preselection can easily be applied so that only the small, relevant part of the database can be downloaded.
  This file is available in the data folder of the project's Github repository

2. **Geodata for each of the neighborhoods in The Hague**

   It is important to get geodata in the form of areas (i.e. outer boundaries of an araa, also known as polygons), rather than points (longitude/ latitude pairs). This way, we can use powerful chloropleth maps.

   This data can be downloaded from the following government website:
   https://data.overheid.nl/dataset/a357969a-0ff8-43db-a3c3-d6c425e53f49

3. **Foursquare venue data for each of the neighborhoods in The Hague**
   After finding the centerpoint for each of the area, we will retrieve any venue contained in location data provider Foursquare's database within a 500 m radius. This yields around 800 venues.

4. **Foursquare ratings data for each venue (where available)**
   Foursquare allows for a deeper look into each of the venues we have retrieved. We are looking for data that says something about the quality of the venue. The best available metric is customer rating. However, there are several problems with this, including:
   - It is subjective
   - It may be biased (only very satisfied & very dissatisfied client submit a rating, as may the owner's friends)
   - There may be a lack of data point (only a few ratings)
   - The price of a product may not be properly taken into account by a customer (e.g. what can you expect from a EUR 5 pizza vs a EUR 20 one?)

   Nonetheless, in the absence of objective quality data, subjective data is all we have, so we have to work with it. But keep its limitations & pitfalls in mind!
   Note that this type of data is considered premium data by Foursquare, so that relatively strict limits on the number of requests per day apply.

   A copy of each of the data set used in this project can be found at:

https://github.com/jevee/Coursera_Capstone/raw/master/data/

METHODOLOGY

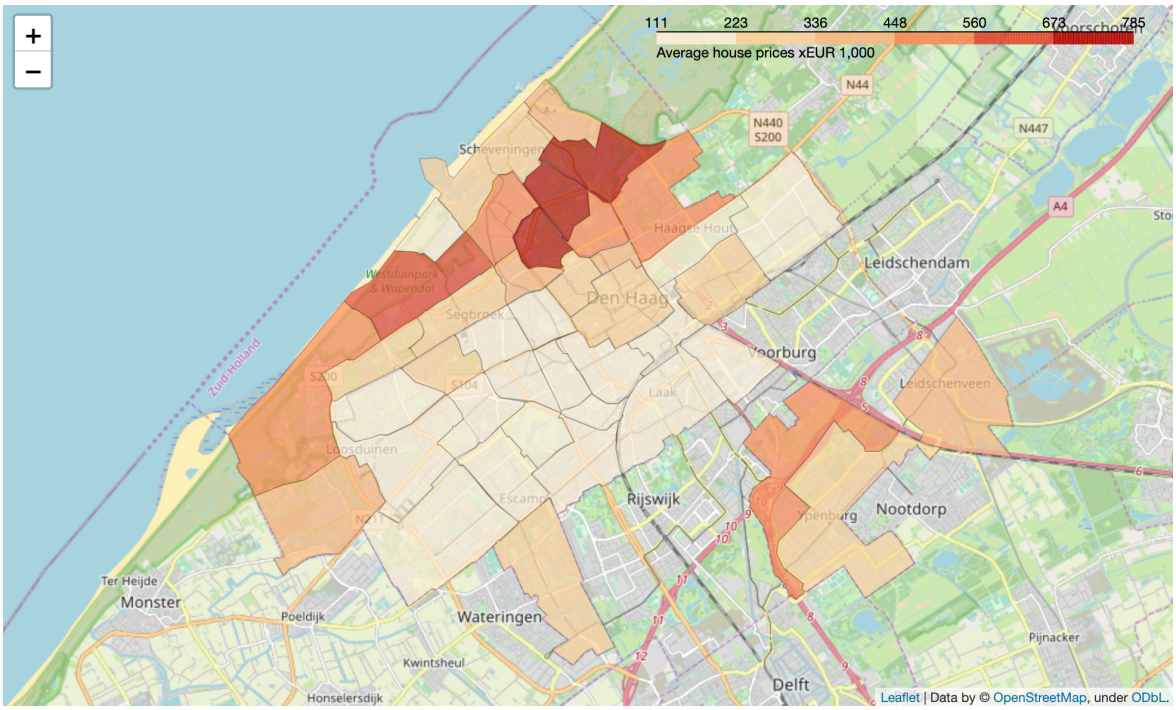
The Dutch Statistics Bureau CBS publishes a wealth of socio-economic information on each neighborhood in The Hague. After assessing the type, quality and completeness of information, it was decided to proceed with 2 key characteristics:

- Average house price – this serves as a proxy for the wealth & income of the neighborhood
- Degree of urbanization – this serves as a proxy for "sophistication". The working assumption is that the more urban the area, the more demanding its residents are when it comes to quality of amenities.

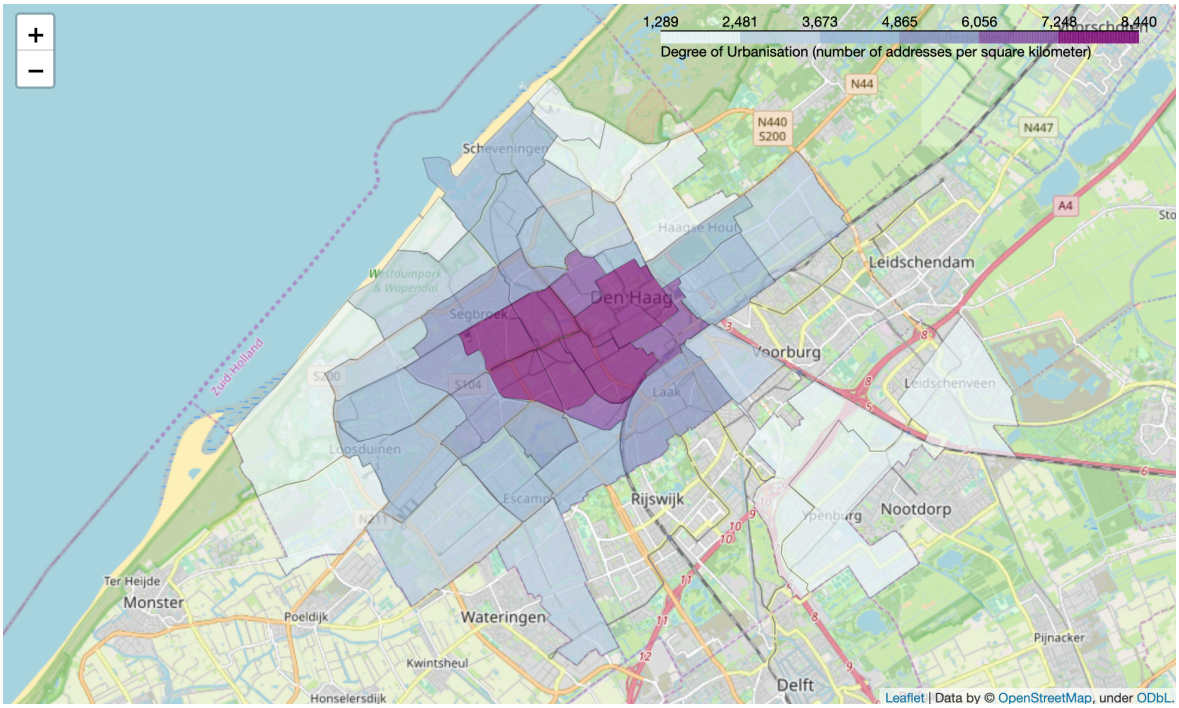
Next up we combine retrieve a json file containing the polygon (area definition) of each of our neighborhoods. We then visualize the variances in house price and degrees of urbanization throughout The Hague using Chloropleth graphs.

These graphs allow us to begin to understand the socio-economic structure of The Hague in an intuitive way. Also, once set-up, we can later add another layer of information on top of these maps.

House price distribution throughout The Hague:



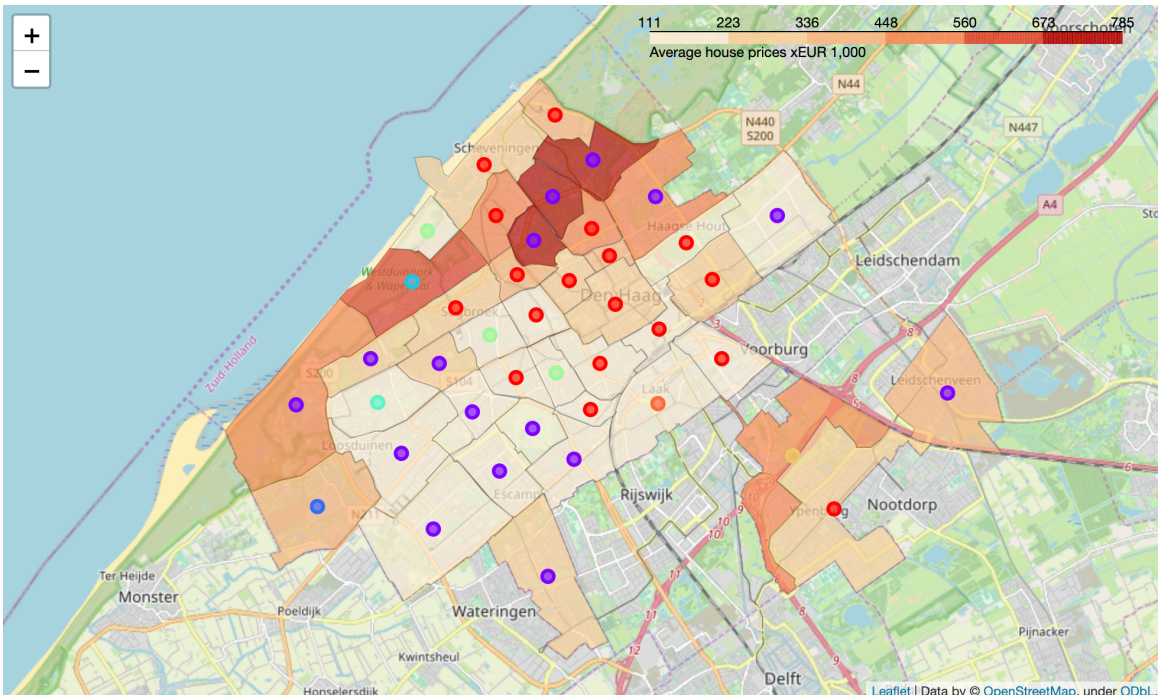
Degree of urbanization throughout The Hague:



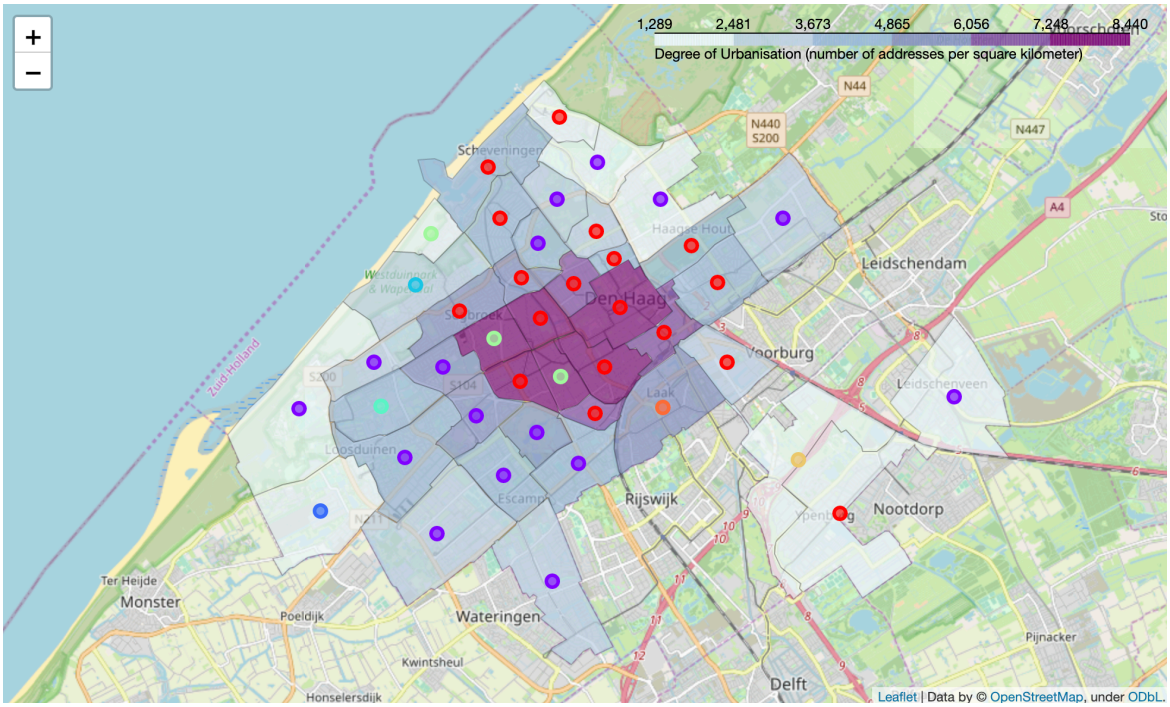
In order to retrieve additional information from Foursquare we first need to find the centerpoint for each of the neighborhoods/ polygons. Once we have that, we can retrieve venues within a certain radius of this centerpoint. This works best if each of the neighborhoods is a perfect circle. Of course, this cannot be the case, but as a proxy, the centerpoint approach is workable.
Finding the centerpoint is very straightforward once the original json data is converted to a Geopandas geodataframe.

We next retrieve all the venues within a 500 meter radius of each of the neighborhoods center points. We will then cluster the city's neighborhoods. The goal of this is to see if there is a relationship between a neighborhoods's wealth or degree of urbanization and the types of venues it has.
If so, this could distort the analysis of venue ratings. For example, lets say that restaurants typically get higher ratings than other types of venues. This means that neighborhood clusters with a high number of restaurants will have a higher perceived quality of amenities. This will then compromise any subsequent analysis on wealth and urbanization.
We will determine this by superimposing the clusters on the Chloropleth maps.
It turns out that there seems no relationship between neighborhood wealth and clustering, but there may be one for degree of urbanization.

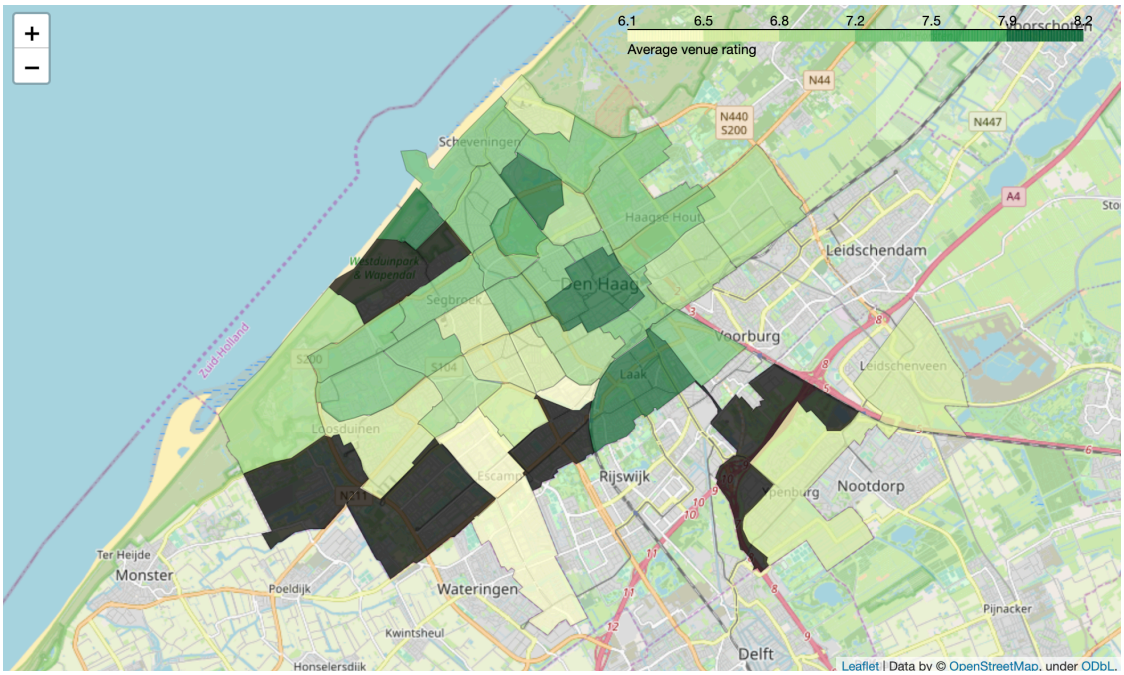
Clusters superimposed on house price distribution:



Clusters superimposed on degree of urbanisation:

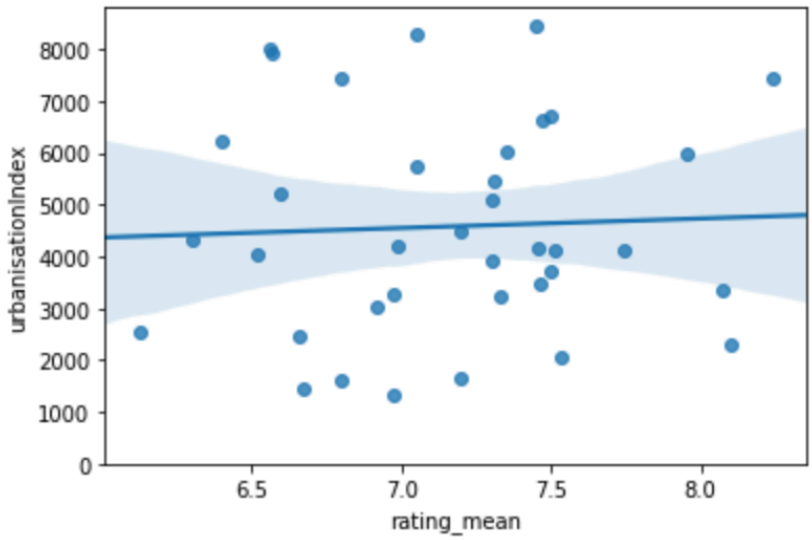


Next, we retrieve the venue detail information for each of the venues from Foursquare. This contains a wealth of information, yet we only use ratings and discard the rest. There is a substantial difference in the average venue quality across the neighborghoods in The Hague:

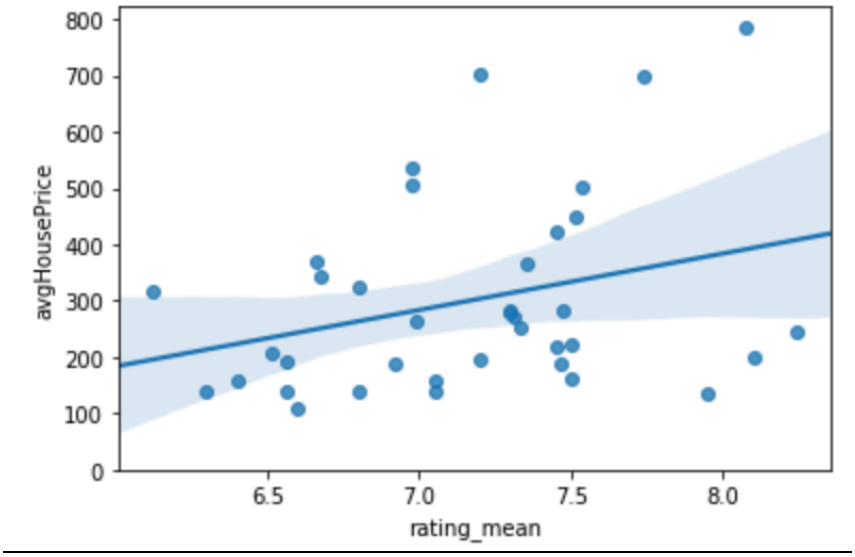


RESULTS & CONCLUSION

A regression of average venue rating against the degree of urbanization shows an almost completely random distribution:



For the average house price, there does appears to be a modest positive correlation:



The Pearson Correlation Coefficient is 0.30 with a P-value of 0.07
This means there could well be a correlation between neighborhood wealth and the (perceived) quality of amenities, yet at best a weak one.