# Capstone Datascience Project

**Introduction**:

Investing in property has been a lucrative business over the past years. The average house prices in the, densely populated, Netherlands have been skyrocketing over the past years. An estimate done by several research institutions showed that between 2017 and 2019, the prices increased on average with 9.7%. This research will focus on the city Delft. “Why Delft?”, it is a beautiful city in the center of the Netherlands, with a strong technical university. This results in a population that is relatively young and therefore with many first-time buyers. This study will develop a model to determine good neighborhoods in Delft to settle for young adults and potentially have children.

**Data Requirements:**

In order to come up with a model that can predict if a neighborhood is suitable for a new family. We need to determine what is important for a new family. After a deep dive( and talking with the subject matter experts, aka. The author, cause I am in that situation), the following features have been selected:

• Vegetation (Trees)

• Water (canals and ponds)

• Playgrounds

• Venues (incl. object of interest)

Based on these features, we can develop a model and evaluate the attractiveness of a neighborhood in which we want to invest. The data should contain at least:

• Longitude and latitude, in order to decide in which neighborhood, they belong.

• Count per neighborhood of the features discussed above.

• Open source, and freely accessible

• We need GEO data on the city of Delft.

**Data Collection:**

The above features were selected, among others, based on their availability. The municipality of Delft provides several open access data sets. These sources will cover: “Vegetation, Water, Playground”. The Four Square API will be used to collect venues per neighborhood.

**Data understanding:**

Before we can prepare the data, we need to study the different data sources:

* Vegetation dataset: list of trees which are taken care of by the municipality, additional information includes longitude, latitude, height, shape, level of maintenance, type, age etc.
* Water dataset: list of water sources such as locations of canals and ponds, the level of maintenance, if it needs mowing and the caretaker etc.
* Playgrounds dataset: playgrounds with longitude, latitude, size, level of maintenance, caretaker and year of development etc.
* Foursquare Api; Venues or object of interest within a certain radius. Additional information includes categories and names of the venues.

The GEOdata per neighborhood is not available and has to be created from scratch. This is done using the tool: <http://geojson.io/>. With this tool, you can draw polygon on the map and later plot these polygons over the folium map.

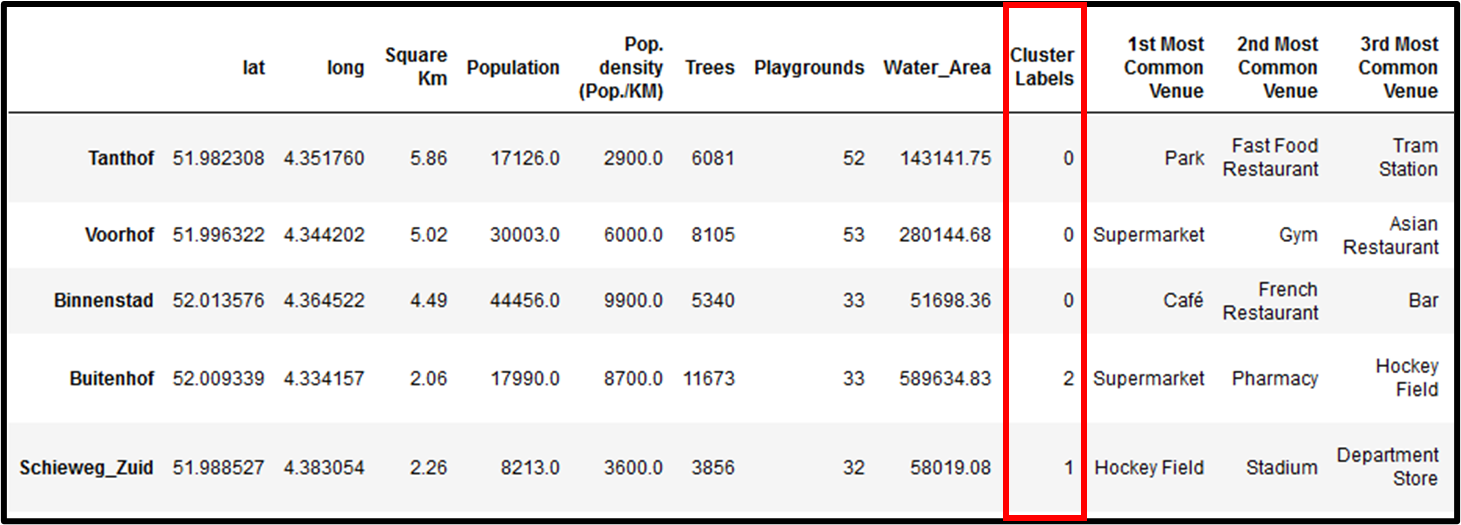
**Data Preparation:**

The opensource data needs little preparation they are already provided in easy handling formats such as csv and Json.

**Clustering venues:**

The below shown data frame is the result of the data collection and preparation steps described in the previous section. The next step is to cluster the venues in order to describe similarities between then. This can then be used to classify them, and allocate weights.

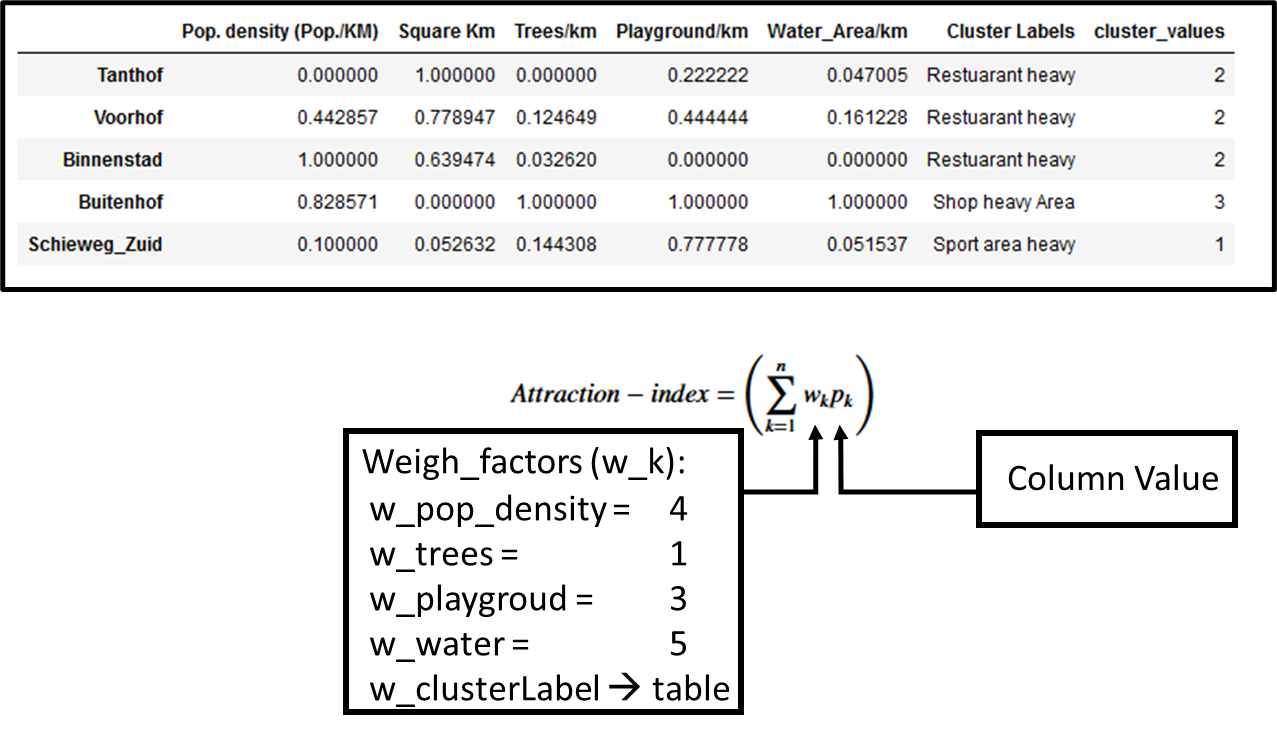
Table 1: Dataframe of the collected and cluster data



**Model development and Results:**

The final dataframe that is used as input for the model is, shown below. Here we added weights to the clusters after the authors preference. Before the column values are used to evaluate the model, they are converted in densities, meaning the column values are divided by the area of the neighborhood. Some areas are larger then others, and this compensates for that. The final model is the result of the sum of the weighted columns values. See equation below. These values are then again normalized with the max value, resulting in the final outcome.

Table 2: Final dataframe used in Model. below that, final model is shown with weighted factors



Final Output!

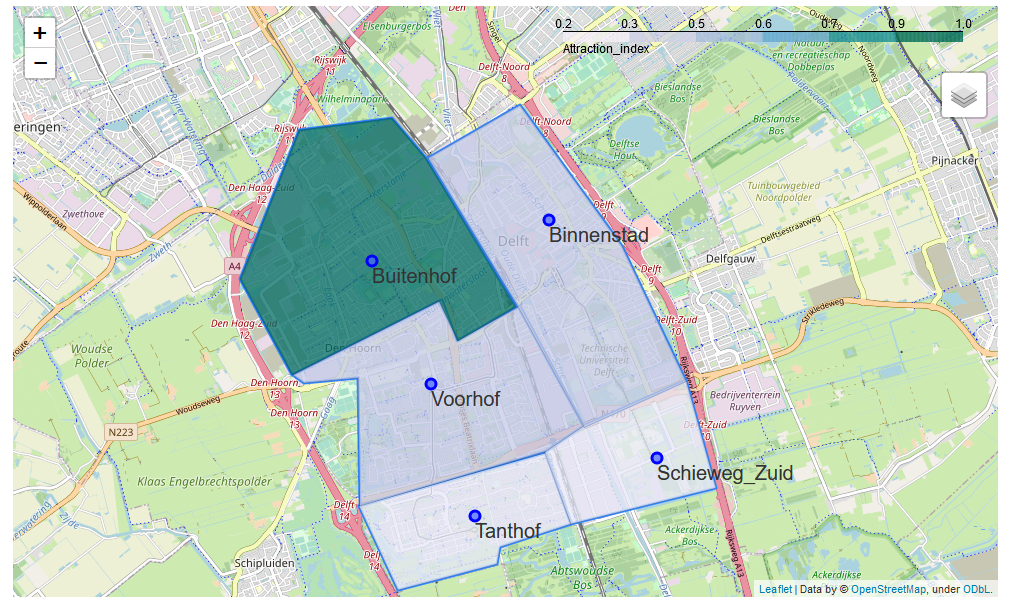
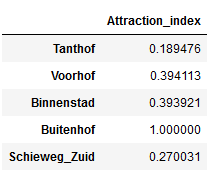


Figure 1: Map of Delft, showing the neighborhoods with color coding for the attraction factor calculated using the model described in Table 2

Table 3: Attraction Index table per neighborhood, as shown on the map in figure 1.



**Discussion & conclusion**

The current model is heavily subjected to the author’s preferences, and can not be used as a general case. Different people value different things. The weight factors should be personalized every time you run the model. The same holds for the feature selection, the current selected features show a strong personal opinion. Many more features can be though of e.g. crime rates, house values, etc. Due to limitation in data availability, these have not been selected. Best place to live in Delft (according to this model) **“Buitenhof”!**