Investigation on Population Migration within Cologne

# Introduction

Cologne is the 4th biggest city in Germany. It is known for it’s very individual culture and people. The city seems to be very heterogeneous in different aspects.   
In recent years, the costs of living and rents have increased by a lot all over Germany, especially in bigger cities. This is of course also true for Cologne.

As a consequence, people are forced to move in these days. In this report, we want to investigate this migration.

## Problem/Question

Moving is always a big decision which you will only do after a lot of thinking, because it’s a big step personally, socially and also financially. We want to understand what makes people leaving their place. We want to find out if there are groups of people who are more probable to move than others.

Also, we want to identify growing and shrinking districts within Cologne. For this we take as a measure the net migration *within* Cologne. People from outside moving to Cologne are neglected in this analysis. After analyzing the population in those districts, we want to attempt to understand why people are coming/leaving.

A stakeholder of this project could be the city of Cologne itself. New investments in housing spaces, park areas and venue settlement support shall be done in the future. The administration wants to better understand the flux of people among city districts to base investments in certain districts on the results of this project. If it can be explained what characteristics of districts make people leave those places, maybe the administration can invest in counter-measures to compensate for the disadvantages. In a similar way, non-growing districts may be made more attractive.

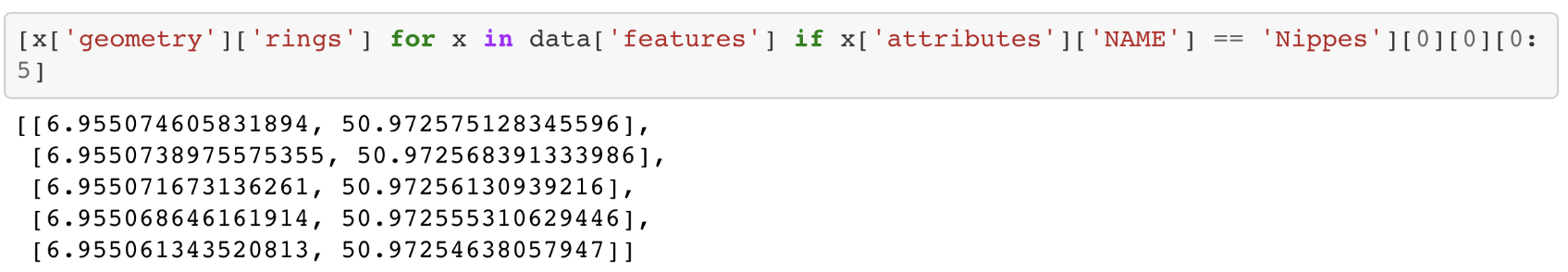
## Data

The city of Cologne has a portal for open data which is located at <https://www.offenedaten-koeln.de>.

For this analysis, I will use some of the data files provided there. Most of the files are given in geojson format. This means, that data is given in an array of features. Each item in the feature list corresponds to a district in Cologne and has attributes and geometries. The attributes dictionary contains the data which depends on the file. The geometry dictionary contains a list of points which correspond to the edges of the wrapping polygon of that district on a map.

For example, we will use a geojson file with polygon geometries for each district. This file has the following attributes:

Where the key ‘NAME’ corresponds to the district’s name. The following districts are available:

Apart from the attributes, each element of the data[‘features’] array has also geometries, here are the first 5 of the district ‘Nippes’:

These points are the edge points of the wrapping polygon of the district.

Each of the geojson files has a similar structure. We find always an array of ‘features’. Each element contains polygon geometries and an attributes dictionary.   
The attributes dictionary has some overhead elements like the name, the id or the number of the district. In addition we find different attribute keys for the different files containing information about different aspects, e.g.

* Net migration for a district
* Ratio of water/built-up/traffic/… area in a district
* Ratio of people living in 1/2/3/4/5 person households
* …

In the notebook these attributes will be extracted for obtaining most of the characteristics of the districts. The only exception is data about the venue environment. In order to obtain a data set for venue locations of a district, Foursquare requests have to be assembled with the center location of each district in latitude/longitude coordinates. Unfortunately, I couldn’t find the center locations of the individual districts. Therefore, I used the polygon geometry data and wrote a simple function to compute the centroid of each polygon. This should give the center location of a district and was used for the requests to the Foursquare API. Foursquare then returns venue data within a certain radius of the location. For details, please refer to the notebook.

In the table below, the used data sources are listed.

I will use data for the following data files:

|  |  |  |  |
| --- | --- | --- | --- |
| Description | Filename | Source |  |
| Districts Location | bezirke.json | <https://www.offenedaten-koeln.de/dataset/stadtbezirke> | Used for extracting the center of each district |
| Inhabitants | einwohner.json | <https://www.offenedaten-koeln.de/dataset/einwohner-statistik-koeln> | Used for finding growing or shrinking districts |
| Area Usage | flaeche.json | <https://offenedaten-koeln.de/dataset/flaechennutzung-koeln-anzahl> | Used for finding area usages in districts like water/park/… |
| Age Distribution | altersgruppen.csv | <https://offenedaten-koeln.de/dataset/einwohner-nach-altersgruppen> | Distribution of age distribution |
| Household distribution | haushalte.json | <https://offenedaten-koeln.de/dataset/haushaltsgr%C3%B6%C3%9Fe> | Distribution of household size |

I will also use the Foursquare API to retrieve venues close to the districts’ center points.