

Verifying Berlin neighborhoods

1. Data acquisition and selection

1.1. Data sources

The first thing which is required is a list of the Berlins neighborhoods. This can be retrieved by the following Wikipage: https://de.wikipedia.org/wiki/Verwaltungsgliederung_Berlins

In the middle of this page one can find a table which provides the required information (date: 07/2019).

Extract [number | neighborhood | borough | area km² | inhabitants | inhabitants per km²]:

Nr. ↕	Ortsteil ↕	Bezirk ↕	Fläche (km ²) ↕	Einwohner ^[2] (30. Juni 2019) ↕	Einwohner pro km ² ↕
101	Mitte	Mitte	10,70	101.932	9526
102	Moabit	Mitte	7,72	79.512	10.299
103	Hansaviertel	Mitte	0,53	5.894	11.121
104	Tiergarten	Mitte	5,17	14.753	2854
105	Wedding	Mitte	9,23	86.688	9392
106	Gesundbrunnen	Mitte	6,13	95.393	15.562
201	Friedrichshain	Friedrichshain-Kreuzberg	9,78	134.900	13.793
202	Kreuzberg	Friedrichshain-Kreuzberg	10,40	154.862	14.891
301	Prenzlauer Berg	Pankow	11,00	164.593	14.963
302	Weißensee	Pankow	7,93	53.737	6776
303	Blankenburg	Pankow	6,03	6.865	1138

For judging the current attractiveness of a neighborhood the foursquare dataset is used. There the number and categories of venues being in a radius of 1000m are collected. Therefore it is also required to have the geographic coordinates of the neighborhoods centers. This will be retrieved by using the geolocator package of Python.

1.2. Data cleaning

The plain data set of neighborhoods is already convenient for a direct usage.

So there are 96 neighborhoods in the dataset available having the six columns coming out of the wiki table (excl. index) and including the geographic coordinates.

Additionally the foursquare app provides venue data in terms of categories and number of venues.

In total there are ~2900 venues available in the neighborhoods being mapped to 319 unique categories.

1.3. Feature selection

Out of the neighborhood table the name and inhabitants per km² are used as parameters.

Unrestricted

The other information of the table are not of interest at the moment. Additionally the geo coordinates are required for every neighborhood. The foursquare data can be reduced to venue category and number of venues.

Kept features	Dropped features	Reason for dropping
<ul style="list-style-type: none"> - Neighborhood name - Inhabitants per km² - Borough name - Latitude - Longitude 	<ul style="list-style-type: none"> - Inhabitants - Area 	Those two features are contained in the inhabitants per km ²
<ul style="list-style-type: none"> - Venue category - Number of venues 	<ul style="list-style-type: none"> - Venue name - Venue latitude - Venue longitude 	This study is sufficient by working with the venue categories