Battle of Neighborhoods in Berlin



Source: Wikimedia-Berlin\_Subdivisions.svg

# 1. Introduction

The metropole Berlin is overwhelming, fascinating and a growing city in middle-east Germany. Until 2030, the population is supposed to grow around 7.5%, predominantly because of inflowing migration. Before moving to a new place, in particular in Berlin, people explore new places in several respects. Being a metropole in Germany, Berlin consists of twelve districts with 96 neighborhoods. All of them share commonalities and have regional distinctions. That requires a systematic neighborhood analysis of various factors. By considering multiple and heterogeneous factors, this analysis supports the decision of people, e.g. as retiree or student, who move to a new place in Berlin.

The project demonstrates how instruments of data science support a systematic neighborhood analysis by processing and jointly analyzing heterogeneous data sources. The respective notebook provides a structured data analysis in order to make the informed decision where to move in Berlin. The analysis consists of three main steps:

First, at the central coordinates of zipcodes in Berlin, more than 13,000 surrounding venues are collected for a systematic assessment. The venues in a venue category, such as hotel, park, supermarket, allow a ranking of the top ten most common venue categories at each zipcode. This information describes the structures of venues within the boroughs. Therewith, this analysis provides an aggregated view on the structure of boroughs by employing a cluster analysis as part of unsupervised machine learning.

Second, in order to compare boroughs regarding several socio-demographic factors, this study considers a selection of features, like unemployment rate, relative number of physicians, housing prices, and rate of academics. A comparative analysis highlights advantages and disadvantages regarding meaningful features of the boroughs.

Third, "top rated venues" show places in the boroughs that attract the attention of Berlin inhabitants.

These steps support the analysis of neighborhoods regarding personal preferences, extract interesting places and can result in a selection of potential locations. The approach requires the implementation of external data sources that provide the features, needs a comparative analysis of these features as well the visualization of geographic information in maps.

# 2. Data, methodology and Python libraries

From the view of the data, this project contains various data sources. Some stem from official statistics of the twelve administrative districts (boroughs) of Berlin on the level of borough or from private statistics of agents of the housing markets and the RESTful APIs of Foursquare or Yelp. Unfortunately, some data is not disaggregated on the level of neighborhoods or addresses. To incorporate all information into a common data basis, the level of boroughs is the best choice, as some data values are rare.

As points of measurement, coordinates of zipcodes represent not overlapping locations within the boroughs. In order to get an in-deep understanding regarding the boroughs, this project utilizes the central coordinates of zipcode to request diverse information. At those points, information about requested venues represent the structure of the city. The respective data sources are:

First, the location data consists of the boroughs, their zipcodes, and coordinates (see: http://www.statistik-berlin-brandenburg.de/produkte/verzeichnisse/zuordnungderbezirkezupostleitzahlen.xls and https://raw.githubusercontent.com/TrustChainEG/postal-codes-json-xml-csv/master/data/DE/zipcodes.de.csv). Zipcodes serve as the index and refer to coordinates, borough names, etc. Further, cohesive location data describes the shape of the boroughs (see: https://raw.githubusercontent.com/m-hoerz/berlin-shapes/master/berliner-bezirke.geojson).

Second, the Foursquare API provides a database of more than 100 million places, globally. We explore the areas around the collected zipcodes in Berlin. Therewith, we perform location search and gather the 100 most famous venues within a circle of 2000 meters radius for each zipcode. The parameters “radius” and “number of venues” are reasonable choices for the concentration for parts of the boroughs.

Third, features of the boroughs are collected by hand from several sources such as official statistics (see: http://www.gsi-berlin.info/gsi\_struktur.asp?kategorie=Sozialdaten, http://www.gsi-berlin.info/gsi\_struktur.asp?kategorie=Gesundheitsdaten: tables: "TG0800221114201859", "TG0800214214201859", "TG0800088114201831-1" ) or private renting agencies (see: https://www.wohnungsboerse.net/immobilienpreise-Berlin/2825).

Fourth, the Yelp API provides data concerning top-rated venues at given coordinates. This information might interest people who consider moving to specific places in Berlin.

From the methodical point of view, this project requires the collection of various data sources from web APIs (Foursquare and Yelp) for the venues and a collection of socio-demographic feature from official and private statistics, which provide a clear view concerning the standard of living in boroughs. The analysis of this raw data extensively needs data transformations, aggregations and comparative plots as means of explanatory data analysis. Further, to obtain a systematic view on the structure of the boroughs, a cluster analysis facilitates a comparison of the locations. Therefore, the venues and their categories were collected at each zipcode in order to compare the relative frequencies of venues per category at each zipcode. These frequencies of venues per category, called "category feature" serve as a measurement of dissimilarity of distinct locations. The cluster analysis groups locations with similar "category features" into a cluster and separates locations with more diverse features. A dendrogram shows the distances between the "category features" in order to determine a plausible number of clusters. Therewith, a hierarchical cluster algorithm provides the cluster labels for the zipcodes. These derivers clusters of similar locations within boroughs of Berlin.

Equipped with these data and tools, one can select some interesting locations. In order explore the selected location in more detail we present top-rated venues at given zipcodes. An example demonstrates the application of this analysis of neighborhoods.

The analysis mainly applies the following Python libraries:

- Pandas, Numpy – Libraries for data storage, manipulation and array computing

- Scipy – Library for dendrogram and hierarchical cluster analysis

- Matplotlib, Folium – Libraries for representing numeric and locational data

- Geopy – Library to retrieve locational data

- Json – Library to handle JSON files

- Requests, Urllib – Libraries to retrieve data and handle http exchange with the Foursquare API and Yelp API