Analysis Of Land Prices And Venues In Berlin

Capstone Project

Alexander Strom

22 March 2020

Contents

1. Background and problem description 3

2. Data requirements and data description 3

3. Methodology 4

3.1 Data collection, data wrangling and data cleaning 4

3.2 Exploratory data analysis on boroughs and statistical data of Berlin 4

3.3 Coordinates of Berlin local centers and Foursquare venue data 5

3.4 Data visualization 6

3.5 Clustering: K-Means 7

4. Results 8

5. Discussion 10

6. Conclusions 10

7. Sources 11

1. Background and problem description

Berlin, the capital of Germany with a total population of 3.6 Million and an area of around 890 km2 [1], has recently grown more and more as a popular city in Central Europe accompanied by "skyrocketing" property prices. A study by the property consultancy Knight Frank, mentioned in [2], attributed the increase in property prices to the population growth, the stable economy, relatively low unemployment rates and interests from investors.

Up to the year 2030 the population in Berlin will approximately grow by 4.7% or 177,000 inhabitants. The average age will rise only relatively gently because of the continuing immigration of young people from 42.7 years in 2018 to 42,9 years in 2030. The positive development of the natural changes of population as the ratio between birth-rate and death-rate due to the ongoing increase in birth-rates will remain stable because it is assumed that immigration of young people will continue. Up to the year 2025, it is assumed, that the birth-rate will exceed the death-rate depending on immigration dynamics [3].

As a consequence of the changes, the social and economic dynamics in the city can be expected to rise. The popularity of the city probably will also lead to a growth in tourism. For people with their businesses positioned in the tourism and cultural sector (city tours, trips, museums, restaurants etc.), the state and development of the city structure might be important. This includes the distribution of venues and their popularity in space and time. This knowledige is key to find popular neighbourhoods in order to place specific offers.

However, some (potential) negative effects for the inhabitants of the city include, e. g. increasing housing prices, increasing prices in food places and bars or a higher number of tourists potentially leading to discomfort for some people. Thus, for policy-makers an analysis of venue data might be interesting to understand the city and the population better.

Therefore, the following questions will be addressed in this project:

1. What are characteristic venues in certain boroughs in the city of Berlin?
2. How are the characteristic venues related to housing or land prices?

2. Data requirements and data description

In order to address the problems mentioned above, population, economic and venue data will be collected and analysed. This includes:

- borough geometry data (GeoJSON) [4],
- geographic positions of local centers [5],
- further information on boroughs i. e. population density [6] and land prices [7] and
- venue data from the Foursquare API of defined local centers [8].

The regional authority for statistics of Berlin and Brandenburg does not offer property prices by boroughs. Therefore, land prices will be used as a proxy for the popularity of a borough. In total, there are 12 boroughs in Berlin. In order to assess the venues in more detail, local centers will be defined based on the postal codes.

3. Methodology

3.1 Data collection, data wrangling and data cleaning

The main Python modules, that will be used for working with the data include:

- Request,
- BeautifulSoup,
- Pandas,
- Numpy,
- Geocoder and
- the Foursquare API.

They will be used for collecting and cleaning the data, reorganizing and transforming data including One-Hot-Encoding.

3.2 Exploratory data analysis on boroughs and statistical data of Berlin

Berlin consists of twelve boroughs, that will be represented spatially with a GeoJSON file from and for which statistical on population density (2019) [6] and land prices in 2018 [7] were investigated. With the help of the *Request* and *BeautifulSoup* packages data could be scraped from the websites and transformed to Pandas-Dataframes as shown below.

Table 1: Cleaned and merged data sets on land prices and population density for the 12 boroughs of Berlin

Borough name	Land price in EUR per km2	Population density in 1000 inhabitants per km2
Mitte	2754.92	9.7
Friedrichshain-Kreuzberg	4482.36	14.4
Pankow	1199.56	3.96
Charlottenburg-Wilmersdorf	2724.02	5.29
Spandau	682.56	2.66
Steglitz-Zehlendorf	861.17	3.01
Tempelhof-Schöneberg	1228.79	6.62
Neukölln	494.23	7.34
Treptow-Köpenick	207.98	1.61
Marzahn-Hellersdorf	436.48	4.35
Lichtenberg	711.02	5.57
Reinickendorf	468.65	2.97

Both, land prices and population density show wide ranges. Borough-based land prices from the year 2018 are between 208 (Treptow-Köpenick) and 4,482 (Friedrichshain-Kreuzberg) Euro per square kilometer. Both boroughs also hold the lowest and highest population densities of 1,610 and 14,370 inhabitants per square kilometer, respectively. The relationship between both parameters is shown below.

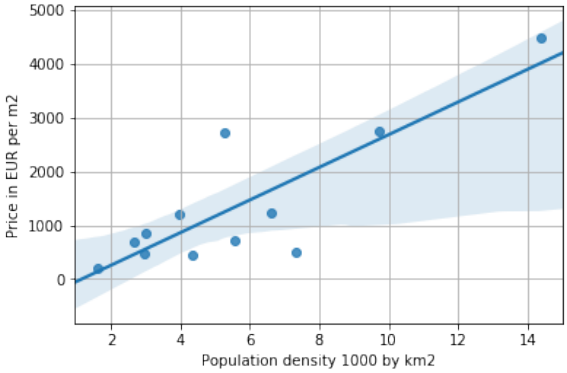

Figure 1: Linear regression between population density and land prices in Berlin

The data suggests a linear relationship between population density of the boroughs and the land prices. The relationshop was expected, but still is an interesting fact. The more people live in a neighbourhood, the higher the land prices will be. One can expect, that this will also affect the category of venues in a local center or neighbourhood.

3.3 Coordinates of Berlin local centers and Foursquare venue data

In total, Berlin has 190 postal codes distributed within the city area. The data is retrieved from [5], which includes all postal codes from Germany. With the selection term „Berlin“, only the postal codes from Berlin will be extracted. The table below shows the results where the „Berlin-column“ was already eliminated. It includes the postal code, latitude and longitude. With these postal codes, which represent the local centers, the venue data from the Foursqure API will be retrieved.

Table 2: Sample data for coordinates of the total 190 local centers/neighbourhoods

Postal code	Latitude	Longitude
10115	52.533707	13.387224
10117	52.518746	13.390193
10119	52.532666	13.407149

The data was retrieved via the Foursquare API in Python. A search radius of 2,500 meters and the maximum limit of 50 venue results was used for each local center. A total of 8,757 venues could be retrieved belonging to 346 unique venue categories. The categories were reduced by assigning primary categories or the first subcategory, because detailed distinction,

e. g. Turkish restaurant/Doner place or History museum/Art museum is not necessary and rather creates noise in our dataset. Below, the last columns show examples of how detailed the subcategories were chosen. By this, the unique categories could be reduced from 346 to 232 categories.

Table 3: Sample data of the Foursquare venue data [7] and the assignment of the primary category

Postal code/ local center	Lon	Lat	Venue name	Venue Lon	Venue Lat	Venue Category	Primary Category
10115	52.533707	13.387224	Dokumentationszentrum | Documentation Center (...	52.535386	13.389668	History Museum	Museum
10115	52.533707	13.387224	Gedenkstätte Berliner Mauer	52.534896	13.390140	Historic Site	Historic Site
10115	52.533707	13.387224	Oslo Kaffebar	52.531029	13.386889	Coffee Shop	Coffee Shop

For further analysis, One-Hot-Encoding was applied to prepare the data for clustering. This means that the categories will be transformed to columns and each row (venue) gets a „1“ or „True“ entry in the respective column. Finally, the data will be aggregated on the local center level, with the mean for each column (primary venue category) representing the frequency of occurrence in the neighbourhood or local center within a radius of 2,500 meters. An example is shown below.

Table 4: Sample data of the preparation of venue data for clustering

Local center	African Restaurant	Airport	American Restaurant	...	Turkish Restaurant	Vegetarian / Vegan Restaurant	Video Store	Waterfront	Windmill	Zoo
10115	0.0	0.0	0.00	...	0.000000	0.03	0.0	0.00	0.0	0.0
10117	0.0	0.0	0.00	...	0.000000	0.02	0.0	0.01	0.0	0.0
10119	0.0	0.0	0.01	...	0.000000	0.02	0.0	0.00	0.0	0.0

3.4 Data visualization

Data visualization is important for understanding (spatial) relationships between data results and communicating these results in order to reach clear conclusions of the analyses. *Matplotlib*, *Seaborn* and *Folium* will be used as the primary visualization tools for

1. understanding the data and relationships between datasets (see section 3.2 Exploratory data analysis) and
2. linking model results (e. g. clusters) with spatial data

by generating bar plots of cluster categories or by creating overview maps of the spatial distribution of the clusters and choropleth maps of land prices.

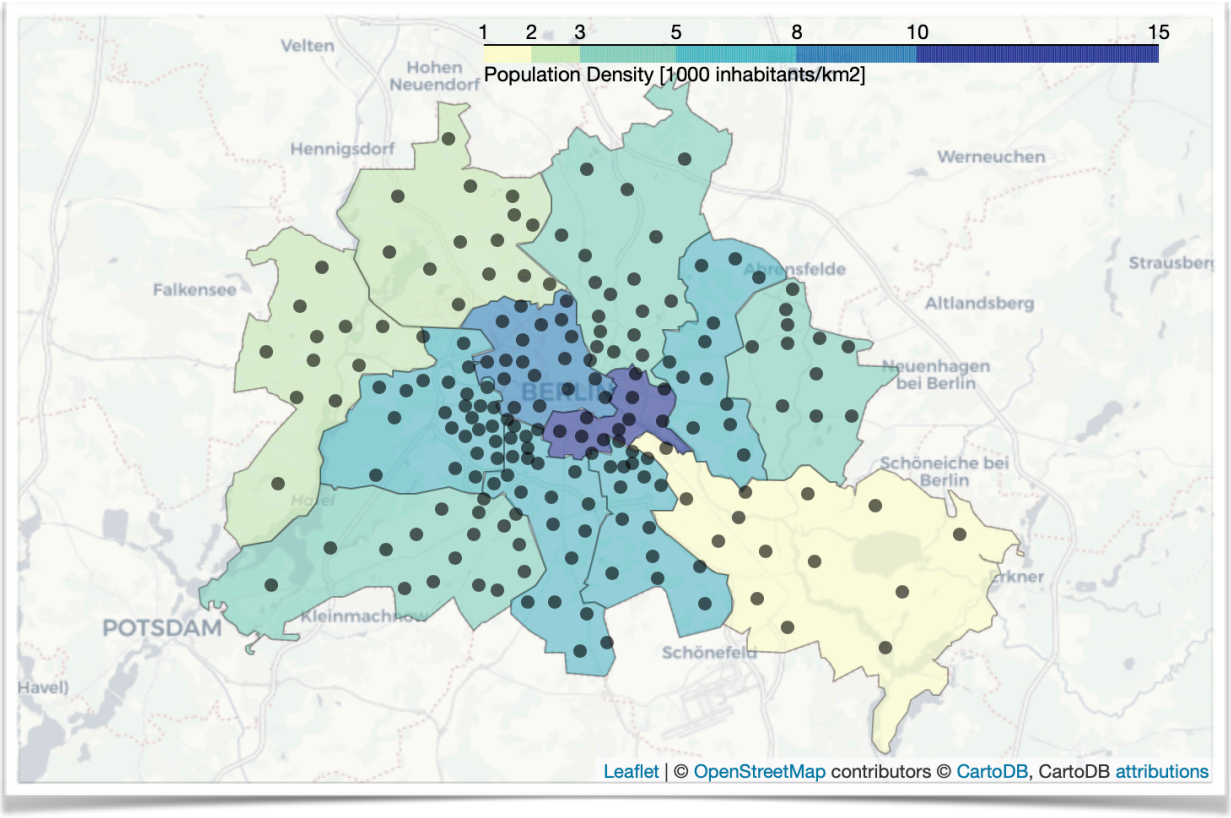

Figure 2: Overview map with locations of local centers, that will be used for clustering and choropleths of population density

3.5 Clustering: K-Means

For further data analysis and modelling, K-Means, a non-supervised machine learning algorithm, will be used for clustering the venue data. For this, the readily available algorithm of the Scikit-Learn package will be applied. Euclidian distance will be used as a metric for the distance between individual venue data points and the centroids defining the clusters. An important unknown will be the number of clusters, that has to be defined beforehand. As the total error (distances of data points to centroids) will always decrease with an increasing number of clusters, the "Elbow method" will be used in order to find the optimal k. The "elbow point" is defined by a sudden softer decrease of the total error with increasing numbers of clusters. Finally, five clusters will be chosen in order to perform clustering.

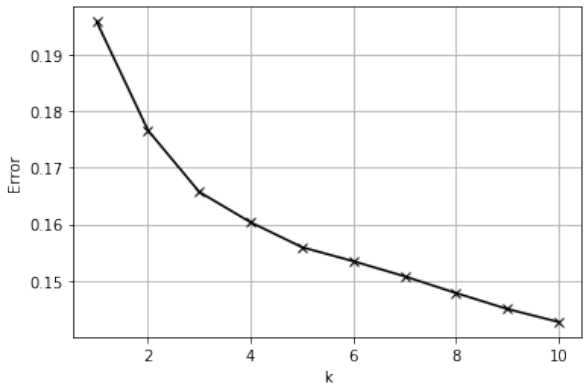

Figure 3: Number of clusters k vs. error to find the optimum k with the „elbow“ method

4. Results

At first, the characteristics of the resulting clusters will be examined and characterized with the help of the frequency of venue categories shown below.

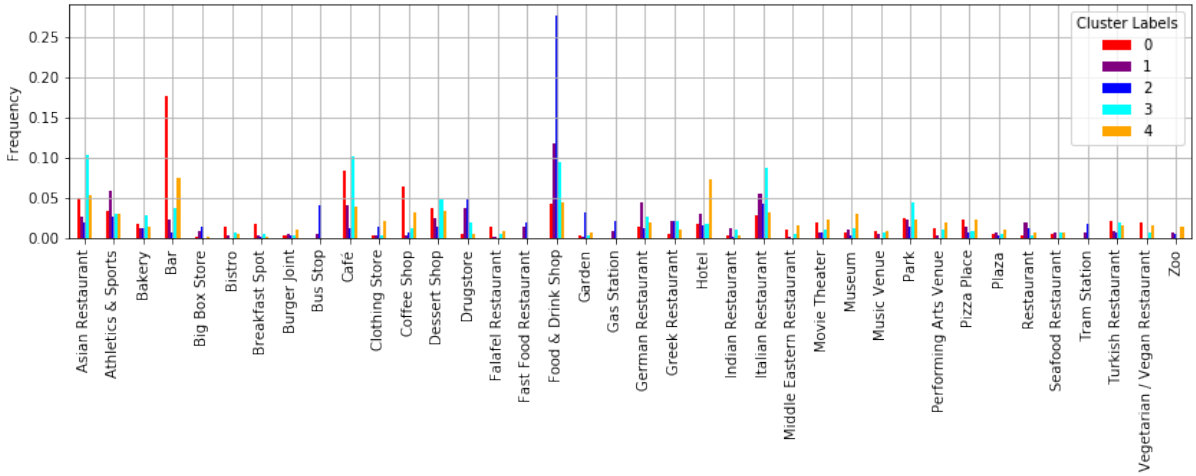

Figure 4: Most popular venue categories and the frequency of occurrence in the clusters

Cluster 0 (red) is characterized mainly by bars, coffee shops and cafés. It also has the highest frequency of vegetarian and vegan restaurants.
Cluster 1 (purple) is dominated by food/drink shops and different types of restaurants.
Cluster 2 (blue) shows a peak for the category „Food & Drink Shops“. It is also affected by bus stops and gardens and has the least bars and cafés. An interesting detail is that it has the highest frequency of tram stations, because they are known to be rather abundant in the Eastern part of Berlin due to the infrastructure in the former German Democratic Republic.
Cluster 3 (green-cyan) is characterized by asian and italian restaurants as well as cafés.
Cluster 4 (orange) is mainly influenced by hotels, bars and some restaurants as well as museums and places related to the category „Zoo“.
The following maps will illustrate the spatial distribution of the cluster categories and their relations to the population density and land prices.

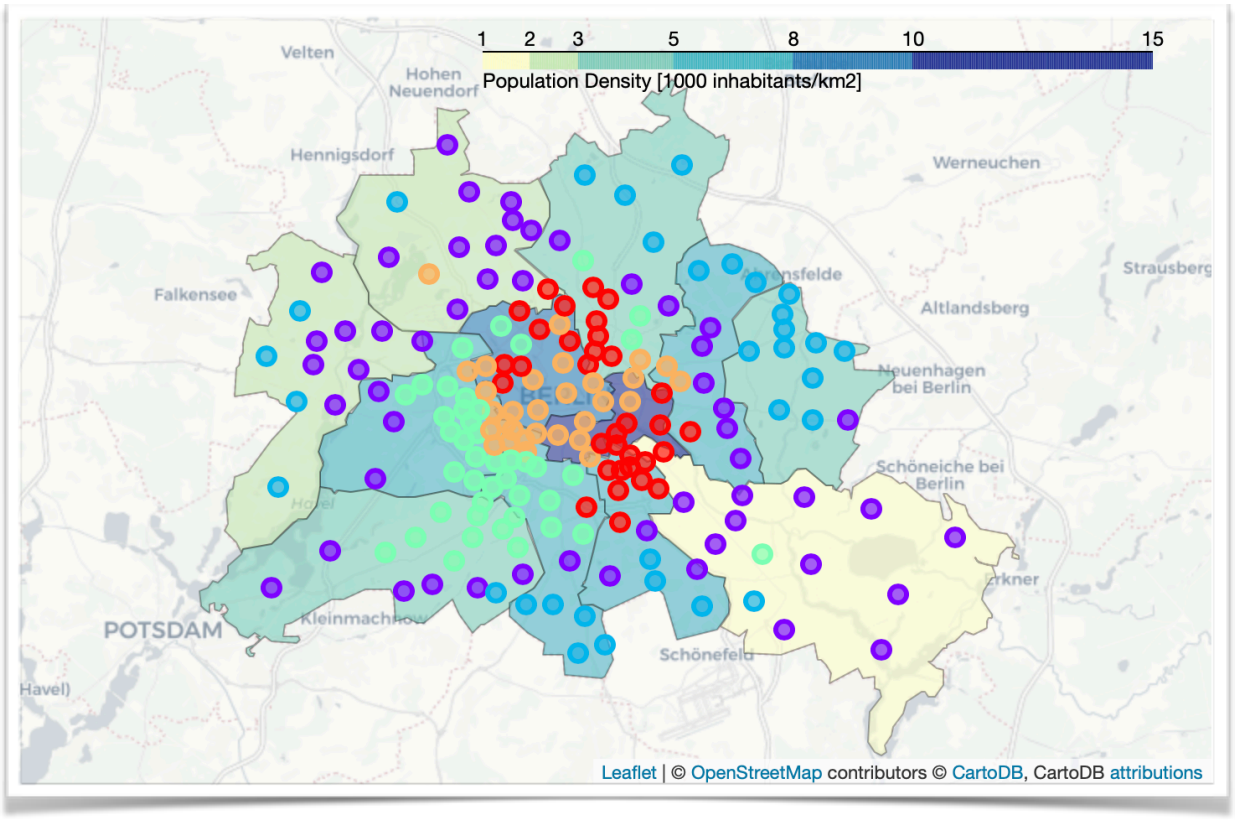

Figure 5: Cluster categories and choropleths of the population density 2019

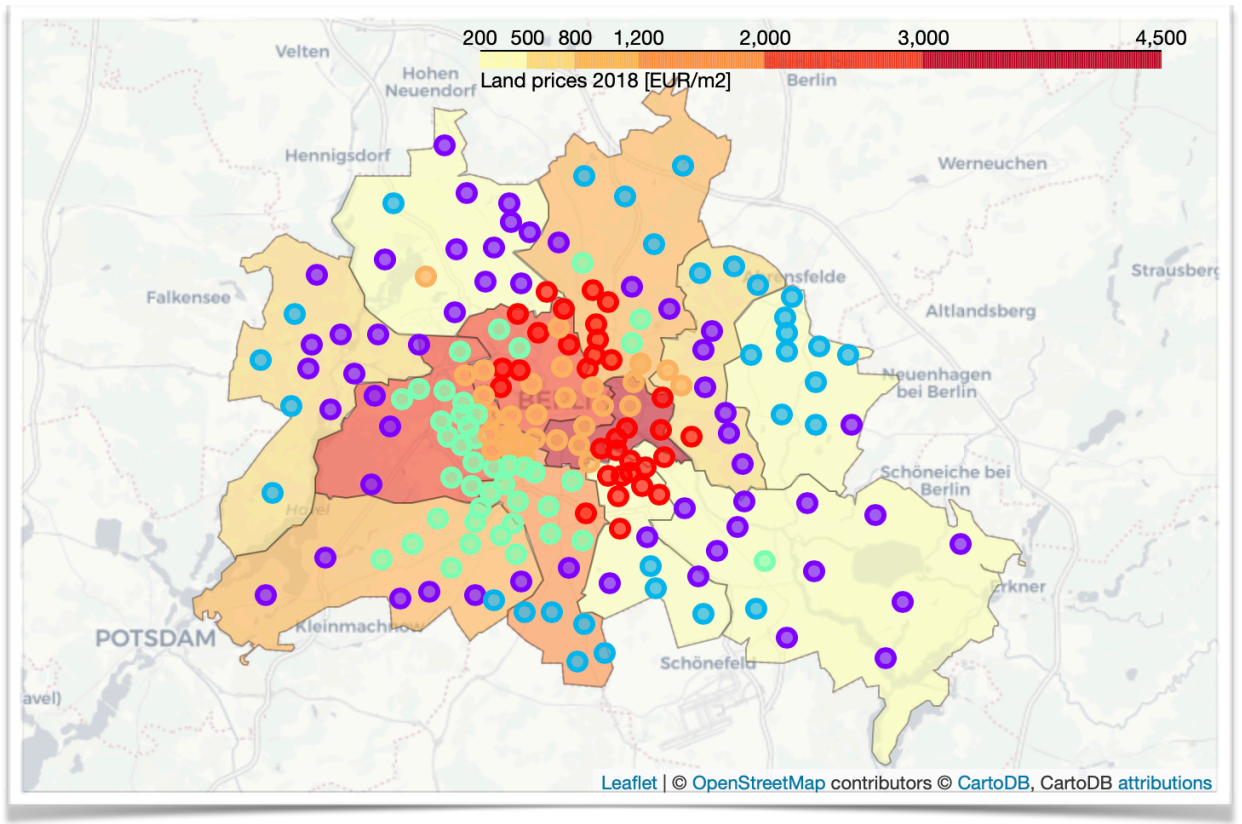

Figure 6: Cluster categories and choropleths of land prices 2018

5. Discussion

The results will be discussed based on the research questions in Chapter 1:

1. What are characteristic venues in certain boroughs in the city of Berlin?
2. How are the characteristic venues related to housing or land prices?

The central part of the city is dominated by cluster categories 0 (red), 3 (green-cyan) and 4 (orange). These clusters are clearly dominated by venues like bars, cafés, restaurants and hotels. Still, there are slight differences, e. g. higher numbers of hotels in one than another cluster or cluster 4 being characterized by the Berlin Zoo. These clusters are linked to higher population densities (5,000 to 15,000 inhabitants per square kilometer) and also higher land prices (2,000 to 4,500 EUR per square meter).

Cluster 2 (blue) shows a low density of cafés and peripheric characteristics, such as bus stops and gardens. Also the population densities and land prices are moderate with 3,000 to 8,000 inhabitants per square kilometer and land prices up to around 2,000 EUR per square meter. This cluster category is dominant in the Eastern part of Berlin, especially in the Northeast. The tram infrastructure is quite strong in those boroughs because of historical reasons. In the German Democratic Republic this type of infrastructure was more popular than the construction of subway/underground tunnels.

Cluster 1 (purple) is spread in the outskirts and is mostly linked to lower population densities and land prices, i. e. 1,000 to 3,000 inhabitants per square kilometer and around 200 to 1,200 EUR per square meter, respectively. Here, the need of touristic infrastructures is very little compared to the boroughs in the city center, which is reflected in the characteristics of the cluster categories.

6. Conclusions

Based on the findings, it can be concluded, that highest land prices (up to 4,500 EUR per square meter) and population densities (up to 15,000 inhabitants per square kilometer) in the boroughs located in the city center of Berlin are linked to venue categories like bars, cafés and hotels reflecting the touristic infrastructure. The population densities and land prices are up to 20 times lower in the outskirts offering good opportunities for people to live cheaply or for business people to open bars and restaurants. For instance, the frequencies of occurrence of coffee places or turkish restaurants are relatively low, so these types of venues might be successful in those neighbourhoods. However, also the population density is low, so the potential of customers is restricted. For every case of potential new venues to be opened, it is recommendable to carry out a more detailed analysis as this project only offers an overview for the whole city.

Based on the results, the following recommendations can be made for policy-makers:

- Make outskirts more attractive for people, e. g. intensify and modernize infrastructure.
- Facilitate the settlement of companies or other economic structures to vivify the boroughs outside the city center.

It could further be shown, that open statistical data from local authorities combined with venue data from the Foursquare API can be powerful tools to analyse local centers or neighbourhoods of cities.

7. Sources

[1] Wikipedia article: https://de.wikipedia.org/wiki/Berlin (accessed: 17 March 2020)

[2] deutschland.de (2018): https://www.deutschland.de/en/topic/life/berlin-property-price-growth-fastest-in-world (accessed: 17 March 2020)

[3] SenStadtWohn (2020): https://www.stadtentwicklung.berlin.de/planen/bevoelkerungsprognose/de/ergebnisse/index.shtml (accessed: 17 March 2020)

[4] Berlin GeoJSON file: https://github.com/funkeinteraktiv/Berlin-Geodaten (accessed: 17 March 2020)

[5] ADFC Fachabteilung Technik (zip codes): 'http://www.fa-technik.adfc.de/code/opengeodb/PLZ.tab' (accessed: 17 March 2020)

[5] Wikipedia data table: https://de.wikipedia.org/wiki/Verwaltungsgliederung_Berlins (accessed: 17 March 2020)

[6] Amt für Statistik Berlin und Brandenburg: https://www.statistik-berlin-brandenburg.de/regionalstatistiken/r-gesamt_neu.asp?Ptyp=410&Sageb=61005&creg=BBB&anzwer=6 (accessed: 17 March 2020)

[7] Foursquare API venue data: https://de.foursquare.com (accessed: 17 March 2020)