Coursera IBM Data Science Professional Certificate
Course 9: Applied Data Science Capstone

Report
Capstone Project: the Battle of Neighborhoods – An Attempt to Cluster Berlin Neighborhoods by Their Living Standards

by Yen-Chun Chen, 2019-06-10

Introduction

Germany's capital Berlin has been known since some years for its gentrification problems. Real estate funds and managers have been most active in the popular areas like Prenzlauer Berg, Friedrichshain or Charlottenburg, where affordable living places became very rare.

It has become difficult not only for tenants to find appropriate housing near their work or their accustomed social environments, but also for regular house owners or investors with less resources to allocate their improvements/investments properly. The objects in the popular areas are now too expensive to be profitable, and the development of the poorer areas is difficult to anticipate. An important aspect of gentrification is that popular areas do not always expand. Ken Steif at urbanspatialanalysis.com calls this phenomenon "supply inelastic" and compares it as the incapability of an area surrounded by water to expand[I].

A Berliner example of a kind of inelastic boundary is the neighborhood Spandau Neustadt situated in the borough Spandau (up to 1920 a city of its own right). It is separated by a huge cross way with a roundabout from the comparably wealthy Spandau Altstadt. Although pertaining to the same locality Spandau, these two neighborhoods haven't been able to melt, until the gentrification of central Berlin compelled the investors to explore Spandau Neustadt because it is actually very well situated (near train, bus, sbahn stations and directly at the river Havel). Yet the question of whether the installation of more luxurious apartments would eventually attract more wealthy residents still remains.

By leveraging the Foursquare location data neighborhoods can be clustered by the types of venues. In a metropolitan area like Berlin, the density of common commerce is, however, not indicative for the area development. The existence of 'special' commodities has to be filtered out to qualify the specific neighborhoods. Using venue data, metrics for measuring the living standards such as 'walkability' or perception of safety[II] could also be implemented in a crude way.

Besides the profitability of an object, the question of the middle to long term development of an area concerns middle class house owners and investors. Obviously, this cannot be answered without historical data. Nevertheless, to extract characteristic 'pictures' of different localities can help to develop indicative factors that will aid one's decision making.

Data

This project required detailed location data, it was acquired by requesting the Foursquare API. For defining the neighborhoods, coordinate data was acquired in different ways, for experimenting with the neighborhood Spandau Neustadt alone, the coordinates could be obtained using browser based Google Maps requests. The coordinates of all Berlin neighborhoods (near to 100) were requested by using the Nominatim class of geopy.geocoders.

The radius of the requests was defined finally at 500m, since one of the main goal here is to experiment with the idea of the 'walkability' of the neighborhoods. The explore endpoint of the Foursquare API returned so-called Points of Interest (POIs). In order to complete the 'picture' of the neighborhoods, extra requests by category Ids for supermarkets, pharmacies, and metro stations were made. Conveniently, the Foursquare data contain the distances between the venues and the coordinates-specified points.

To experiment with the idea of security perception of a certain neighborhoods, venues related to keywords such as 'gamble' or 'sex' were requested, but no results could be obtained. This caused a lamentable fall back of completing neighborhood clusters by their negative aspects[III].

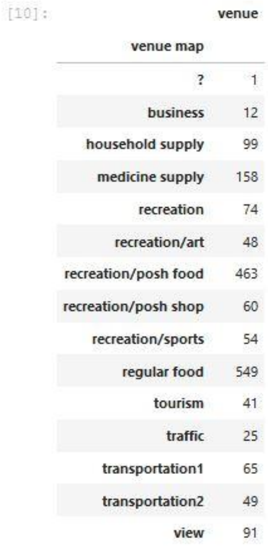
To plot Choropleth maps, a geojson for Berlin's boroughs was found on GitHub[IV].

A minimal csv file containing the average rent price of every borough in 2017 was created, so the results of the clustering could be very roughly validated.

Methodology and exploratory data analysis

The preliminarily defined objectives were 1., to use venues as characteristic features and obtain a 'picture' or 'image' of the Berlin neighborhoods without exploiting users' data. 2., to develop suitable metrics for the measurement of the neighborhoods' current fitness for urban evolution. 3., to cluster the neighborhoods based on a combination of the defined fitness scores and the characteristic features.

The first foursquare venue request within a radius of 3km around all coordinated points returned 6826 data points (POI venues) and 328 unique venue categories were found. These categories are very detailed, they may serve for defining how multicultural or diversified a neighborhood is, but not for the calculation of the urban fitness. In order to increase the discriminating factor of the categories, they were hard coded (no machine learning tool existent for this purpose) to 15 meta-categories:



[10]:

venue map	venue
?	1
business	12
household supply	99
medicine supply	158
recreation	74
recreation/art	48
recreation/posh food	463
recreation/posh shop	60
recreation/sports	54
regular food	549
tourism	41
traffic	25
transportation1	65
transportation2	49
view	91

The tags containing 'recreation\/*.' can be used to filter out more characteristic features, since they can be distinguished from the quotidian necessities such as 'regular food' or 'household supply'. At the same time, they are distinct among themselves.

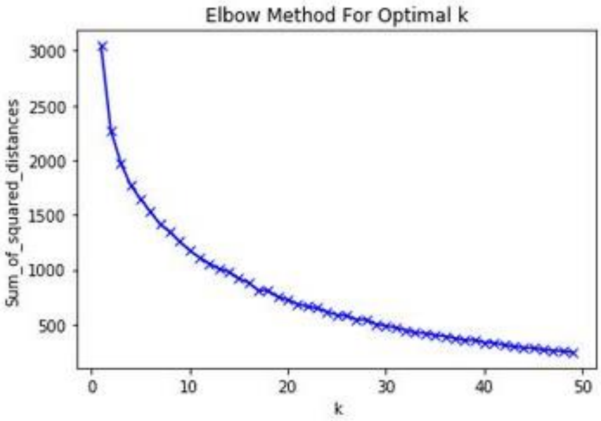
The main effort of this project was dedicated to the reflection on the so-called 'walkability' score. It has become extremely important for city inhabitants that they can reach places of quotidian importance on foot. Although Berlin is very privileged in this aspect, a good selection of daily supply nearby can generally be considered as advantageous.

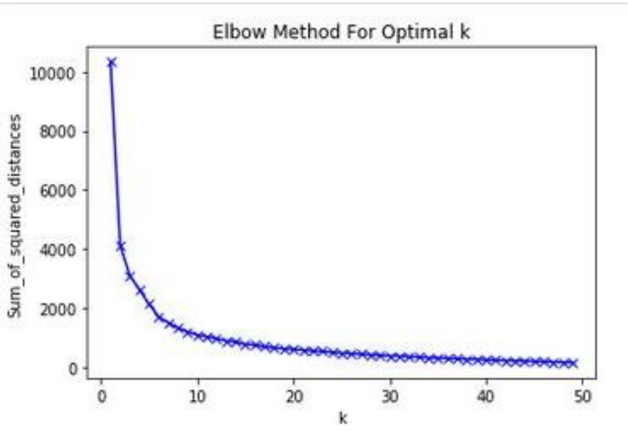
To compute this score, I chose to start with a linear formula, which counts the venues of the categories 'regular food', 'transportation2' (referring to more powerful public transportation facilities like metro stations, light rail stations, train stations and tram stations), 'household supply', 'medicine supply' and 'recreation'. The counts were each multiplied with weights specified as 0.20, 0.30, 0.05, 0.10 and 0.35. Since in a metropolitan area, 'normal' shops are quite evenly distributed, the 'recreation' category, which includes also more specialized offers, was given a heavier weight. The sum of the weighted counts is then divided by the sum of venues counted. The weighting and the meta-categorization can of course be adjusted individually. Mathematically the model can be expressed as:

Category: type of venues within 500 m radius
Score: $S, \quad 0 \leq S \leq 1$
Weight: $W_{i\,,} \; 0 \leq W_i \leq 1, \; i = 1, \dots, N, \; N = number\ of\ categories$
$\sum_{i=1}^{N} Wi \; = \; 1$
Venue: number of venues in each category
$V_j, \quad j = 1, \dots, N$
M: total amount of venues
$M = \sum_{i=1}^{N} V_j$

$$S = \frac{1}{M} \sum_{i=1}^{N} W_i \, V_i$$

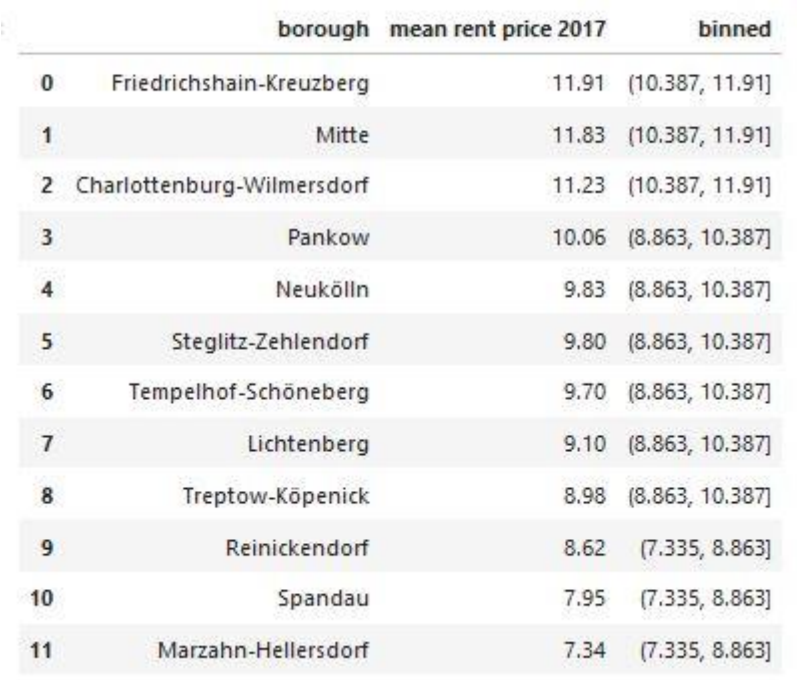
The location data and the simple model described above were used to experiment with the KMeans clustering. The data set was wrangled in different ways to be fed into the KMeans clustering algorithm. For each test, the Elbow method was used to find the best k.





The left figure shows the k/SSE with all original Foursquare categories. The right one shows the k/SSE with meta-categories. It is interesting to see that without refining the categories, the error rate lower much slower.

After every clustering, a Choropleth map was created to get the first impression. The labels of the boroughs were taken by the mean of their neighborhoods label. Finally the clustered results were 'binned' into three sections and compared with one set of real data.



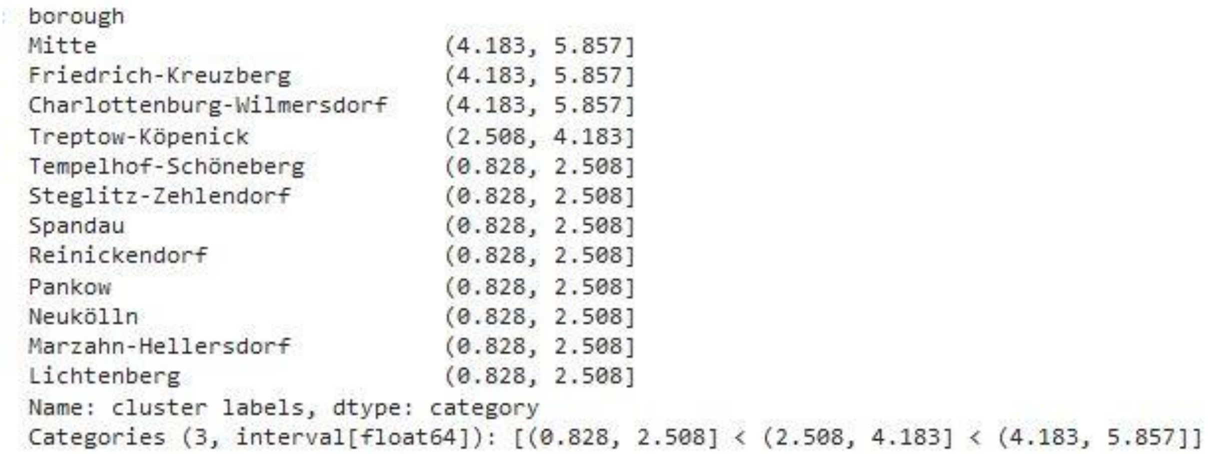
	borough	mean rent price 2017	binned
0	Friedrichshain-Kreuzberg	11.91	(10.387, 11.91]
1	Mitte	11.83	(10.387, 11.91]
2	Charlottenburg-Wilmersdorf	11.23	(10.387, 11.91]
3	Pankow	10.06	(8.863, 10.387]
4	Neukölln	9.83	(8.863, 10.387]
5	Steglitz-Zehlendorf	9.80	(8.863, 10.387]
6	Tempelhof-Schöneberg	9.70	(8.863, 10.387]
7	Lichtenberg	9.10	(8.863, 10.387]
8	Treptow-Köpenick	8.98	(8.863, 10.387]
9	Reinickendorf	8.62	(7.335, 8.863]
10	Spandau	7.95	(7.335, 8.863]
11	Marzahn-Hellersdorf	7.34	(7.335, 8.863]

The 'binning' resembles the idea of clustering. Here the 'bin' parameter was specified at 3.

Results

The result of this ‘walkability’ calculation didn’t match the rent price of Berlin very well, e.g. Schöneberg, which is less expensive than Charlottenburg, had 0.27 while the latter had 0.26. On the other hand, Mitte (0.33), Friedrichshain (0.29), Kreuzberg (0.30) and Prenzlauer Berg (0.28) have all scores approximately in the same range; Spandau Neustadt and Hellersdorf have 0.2 and 0.17 respectively. Furthermore, the poorer Spandau Neustadt differs in 0.02 points from the less poor Spandau.

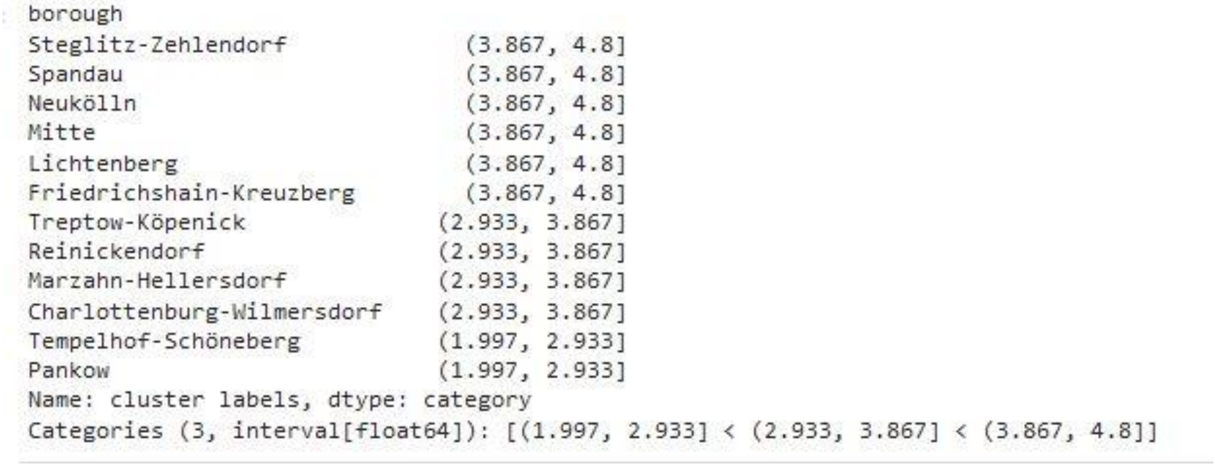
The clustering was done with 1., all original Foursquare categories, 2., all meta-categories and 3., the ‘walkability’ scores combined with other meta-categories. Their results similarly didn’t match in respect to the housing price situations. For example, the first clustering gave the following labels to the boroughs:



```
borough
Mitte                          (4.183, 5.857]
Friedrich-Kreuzberg            (4.183, 5.857]
Charlottenburg-Wilmersdorf     (4.183, 5.857]
Treptow-Köpenick               (2.508, 4.183]
Tempelhof-Schöneberg           (0.828, 2.508]
Steglitz-Zehlendorf            (0.828, 2.508]
Spandau                        (0.828, 2.508]
Reinickendorf                  (0.828, 2.508]
Pankow                         (0.828, 2.508]
Neukölln                       (0.828, 2.508]
Marzahn-Hellersdorf            (0.828, 2.508]
Lichtenberg                    (0.828, 2.508]
Name: cluster labels, dtype: category
Categories (3, interval[float64]): [(0.828, 2.508] < (2.508, 4.183] < (4.183, 5.857]]
```

While Mitte, Friedrichshain-Kreuzberg and Charlottenburg-Wilmersdorf were labeled ‘correctly’[V] as the same group, from Treptow-Köpenick on the boroughs remained undistinguishable. A possible reason could be the high density of similar types of venues in Mitte, Kreuzberg, or Charlottenburg-Wilmersdorf, e.g. art galleries, theaters and concert venues.

The clustering with both ‘walkability’ score and meta-categories didn’t match any better:



```
borough
Steglitz-Zehlendorf               (3.867, 4.8]
Spandau                           (3.867, 4.8]
Neukölln                          (3.867, 4.8]
Mitte                             (3.867, 4.8]
Lichtenberg                       (3.867, 4.8]
Friedrichshain-Kreuzberg          (3.867, 4.8]
Treptow-Köpenick                (2.933, 3.867]
Reinickendorf                   (2.933, 3.867]
Marzahn-Hellersdorf             (2.933, 3.867]
Charlottenburg-Wilmersdorf      (2.933, 3.867]
Tempelhof-Schöneberg            (1.997, 2.933]
Pankow                          (1.997, 2.933]
Name: cluster labels, dtype: category
Categories (3, interval[float64]): [(1.997, 2.933] < (2.933, 3.867] < (3.867, 4.8]]
```

While Pankow and Tempelhof-Schöneberg were labeled quite nicely as belonging to the same group, Charlottenburg-Wilmersdorf was clustered to one of the least popular area Marzahn-Hellersdorf. The linear formula for the 'walkability' score is definitely not sufficient to model human preferences. The lack of the 'negative' venues also contributed to the difficulty. Maybe the KMeans clustering is just not suitable for this kind of tasks?

Discussion

The location data of Foursquare originated by crowd sourcing. Moreover, they tend to refer to 'interesting' places. It reflects, in a certain sense, human behavior. Possibly that's why the unmanipulated data within the spectrum of POIs gave a reasonably good result.

The 'walkability' score was just a small experiment for the beginning. With the current linear formula, the score will always stay quite low, because there is practically no limit defined. Fancier areas, however, tend to have higher scores. During this project, one question always appeared: How can psychology be modeled, i.e. how to model what people prefer, when they are satisfied, where the limit is when more venues are not increasing the happiness. There are many different ways to set a limit, for example with an inverse exponential function.

Yet, even if a much better model can be formulated for a 'walkability' score, there will still be interferences between the score and the clustering algorithm. Does the quality of clustering only depend on the data?

Conclusion

It is in any case difficult to model human behavior. By clustering the neighborhoods of Berlin, I come to the conclusion that places and people influence each other. And in this case, a simple clustering without inference could even be better. For a future project, with more detailed data available, a more comprehensive sampling of geospatial and historical data (e.g. prices, historical location data) could result in fitly description of interactions between people and places.

[I] http://urbanspatialanalysis.com/riding-and-clustering-the-gentrification-wave/; Steif offers here an interesting discussion based on data scientific methods.

[II] see e.g. https://arxiv.org/pdf/1808.02547.pdf

[III] from the perspective of future inhabitants.

[IV] https://github.com/funkeinteraktiv/Berlin-Geodaten.

[V] It is naturally 'incorrect' to say so, since the model hasn't been developed at all.