## *Introduction*

The Berlin S-Bahn has 157 stations and is one of the largest public transit systems in Europe. It is used by 478 Millionen passengers per year.

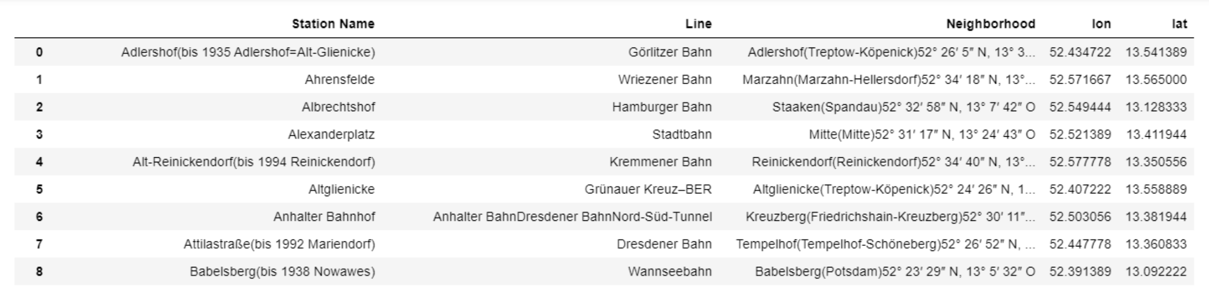
For this project, we want to look at the neighborhoods surrounding S-Bahn stations and classify them. Some neighborhoods are mostly residential, others have more business or commercial spaces surrounding them. The venues closest to a station determine why and how people use it. E.g. if there are no professional places in a neighborhood its residents are likely to travel to other areas for work. This creates daily migrations of people.

By analyzing this data we can classify stations by their primary usage. This data can be useful for city planners to determine where from and where to people are most likely to travel for work and leisure, plan further extension of the network and find places for new development.

## *Data*

We'll need data on the location of stations and on the venues closest to them.

1. List of stations and their geographical coordinates — scraped from this page: https://de.wikipedia.org/wiki/Liste_der_Stationen_der_S-Bahn_Berlin



| | Station Name | Line | Neighborhood | lon | lat |
|---|---|---|---|---|---|
| 0 | Adlershof(bis 1935 Adlershof=Alt-Glienicke) | Görlitzer Bahn | Adlershof(Treptow-Köpenick)52° 26′ 5″ N, 13° 3... | 52.434722 | 13.541389 |
| 1 | Ahrensfelde | Wriezener Bahn | Marzahn(Marzahn-Hellersdorf)52° 34′ 18″ N, 13°... | 52.571667 | 13.565000 |
| 2 | Albrechtshof | Hamburger Bahn | Staaken(Spandau)52° 32′ 58″ N, 13° 7′ 42″ O | 52.549444 | 13.128333 |
| 3 | Alexanderplatz | Stadtbahn | Mitte(Mitte)52° 31′ 17″ N, 13° 24′ 43″ O | 52.521389 | 13.411944 |
| 4 | Alt-Reinickendorf(bis 1994 Reinickendorf) | Kremmener Bahn | Reinickendorf(Reinickendorf)52° 34′ 40″ N, 13°... | 52.577778 | 13.350556 |
| 5 | Altglienicke | Grünauer Kreuz–BER | Altglienicke(Treptow-Köpenick)52° 24′ 26″ N, 1... | 52.407222 | 13.558889 |
| 6 | Anhalter Bahnhof | Anhalter BahnDresdener BahnNord-Süd-Tunnel | Kreuzberg(Friedrichshain-Kreuzberg)52° 30′ 11″... | 52.503056 | 13.381944 |
| 7 | Attilastraße(bis 1992 Mariendorf) | Dresdener Bahn | Tempelhof(Tempelhof-Schöneberg)52° 26′ 52″ N, ... | 52.447778 | 13.360833 |
| 8 | Babelsberg(bis 1938 Nowawes) | Wannseebahn | Babelsberg(Potsdam)52° 23′ 29″ N, 13° 5′ 32″ O | 52.391389 | 13.092222 |

*Cut-Out of the table with the location data of the S-bahn stations, data already cleaned*

Following steps were done to clean the data to get the dataframe that is shown above:

- Deleting of the unnecessary Columns
- Translations of the column names from German to English
- Separating the coordinates from the Neighborhood column by removing all letters
- Seperating the minutes and seconds of the coodinates to transform them to a format that is supported bei thre API
- Transformation of the coordinates from String to float
- Calculation of the longitude and latitude values and deleting of the columns, that were created for the separation and the calculation
- Deleting of all lines that do not contain values eigher in the longitude ir in the latitude

2. Foursquare API to explore venue types surrounding each station. Foursquare outlines these high-level venue categories with more sub-categories.

- Arts & Entertainment (4d4b7104d754a06370d81259)
- College & University (4d4b7105d754a06372d81259)
- Event (4d4b7105d754a06373d81259)
- Food (4d4b7105d754a06374d81259)
- Nightlife Spot (4d4b7105d754a06376d81259)

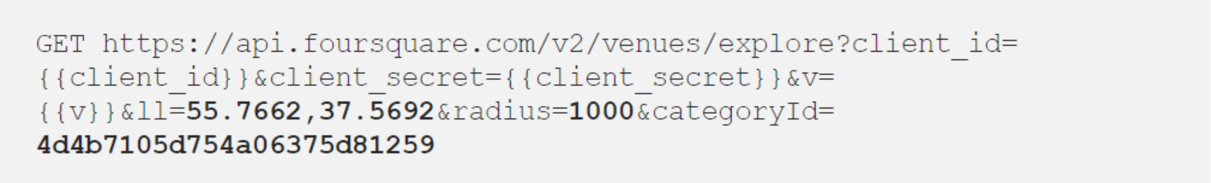
- Outdoors & Recreation (4d4b7105d754a06377d81259)
- Professional & Other Places (4d4b7105d754a06375d81259)
- Residence (4e67e38e036454776db1fb3a)
- Shop & Service (4d4b7105d754a06378d81259)
- Travel & Transport (4d4b7105d754a06379d81259)

We'll be querying the number of venues in each category in a 1000m radius around each station. This radius was chosen because 1000m is a reasonable walking distance.

## Methodology

Exploratory analysis & basic cleanup

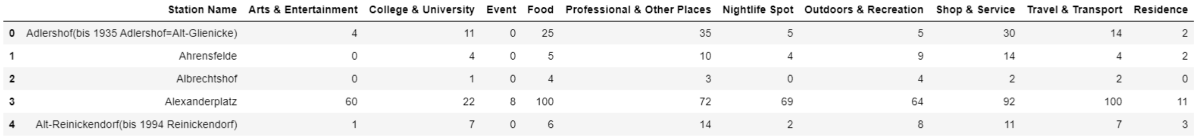
We can use the Foursquare explore API with category ID to query the number of venues of each category in a specific radius. The response contains a totalResults value for the specified coordinates, radius and category. Sample request (1000m radius and category Professional & Other Places):



```
GET https://api.foursquare.com/v2/venues/explore?client_id=
{{client_id}}&client_secret={{client_secret}}&v=
{{v}}&ll=55.7662,37.5692&radius=1000&categoryId=
4d4b7105d754a06375d81259
```

*Example request*

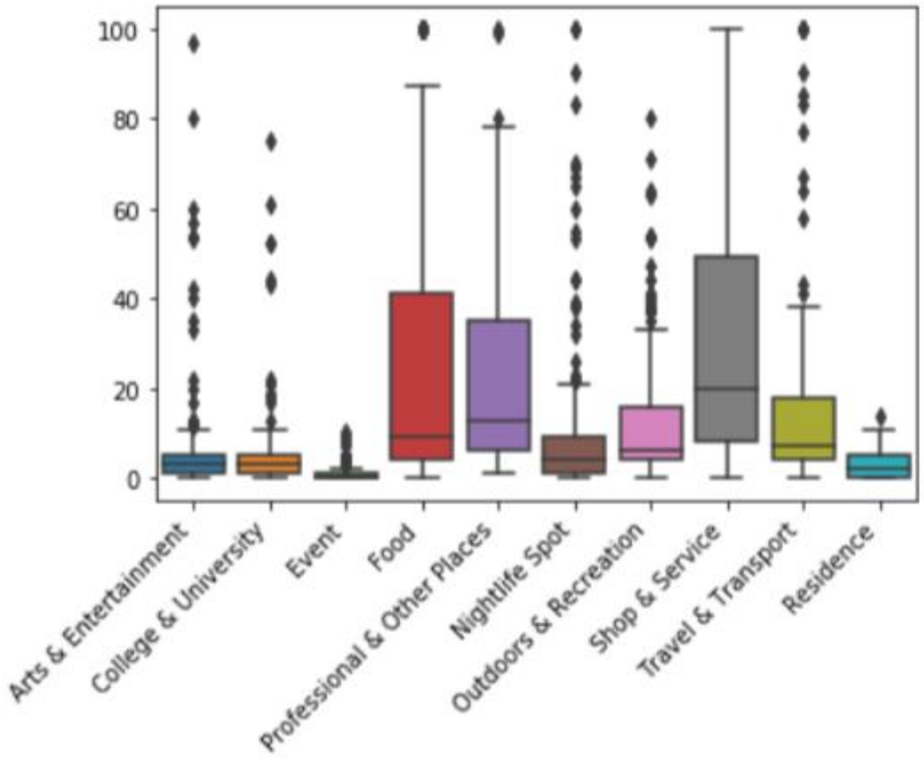
A function was created to run the requests automatically and safe the result in a datafram. Afterwards, dummy columns were created for the venue category column. Finally, the datafram was grouped by the station name. The final datafram contains the number of venues of every category in the neighbourhood of the respective station. The full datafram is available on Github in the notebook.



| | Station Name | Arts & Entertainment | College & University | Event | Food | Professional & Other Places | Nightlife Spot | Outdoors & Recreation | Shop & Service | Travel & Transport | Residence |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Adlershof(bis 1935 Adlershof=Alt-Glienicke) | 4 | 11 | 0 | 25 | 35 | 5 | 5 | 30 | 14 | 2 |
| 1 | Ahrensfelde | 0 | 4 | 0 | 5 | 10 | 4 | 9 | 14 | 4 | 2 |
| 2 | Albrechtshof | 0 | 1 | 0 | 4 | 3 | 0 | 4 | 2 | 2 | 0 |
| 3 | Alexanderplatz | 60 | 22 | 8 | 100 | 72 | 69 | 64 | 92 | 100 | 11 |
| 4 | Alt-Reinickendorf(bis 1994 Reinickendorf) | 1 | 7 | 0 | 6 | 14 | 2 | 8 | 11 | 7 | 3 |

*Datafrom with the numberf venues of every category in the neighbourhood of the respective station*

Let’s display the number of venues as boxplots (showing the average count, spread and outliers).

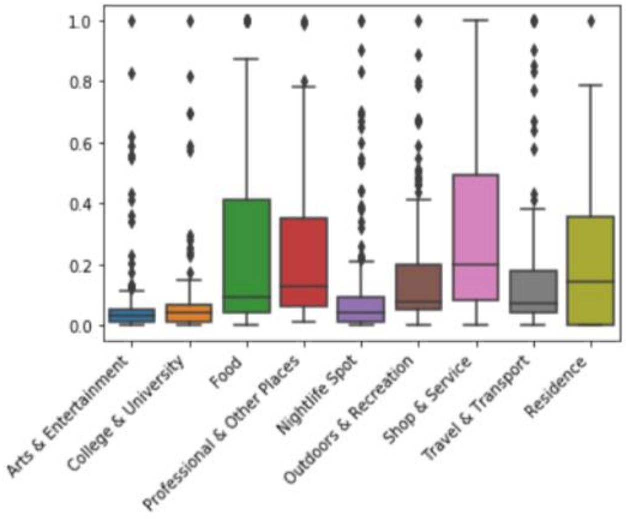


*Boxplots of number of venues in each category*

We can see that the most frequent venue categories are Food, Shop & Service and Professional & Other Places. Event has very little data, so we’ll discard it.

Data preparation

Let’s normalize the data using min-max scaling (scale count of venues from 0 to 1 where 0 is the lowest value in a set and 1 is highest). This both normalizes the data and provides an easy to interpret score at the same time. The scaled diagram looks like this:



*Boxplots of scaled number of venues in each category*

Clustering

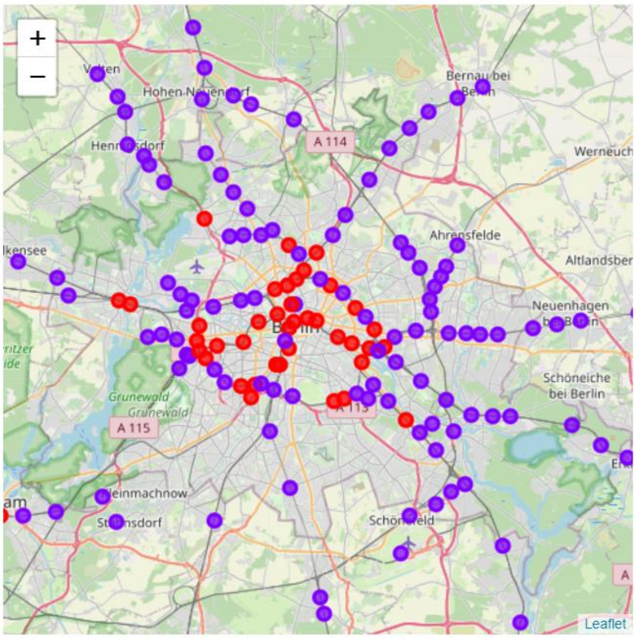
We’ll be using k-means clustering. These were the preliminary results with different number of clusters:

2 clusters: show the uptown/downtown divide

3 clusters: clustering within the downtown, also identify neighborhoods with very low number of venues

## *Results*

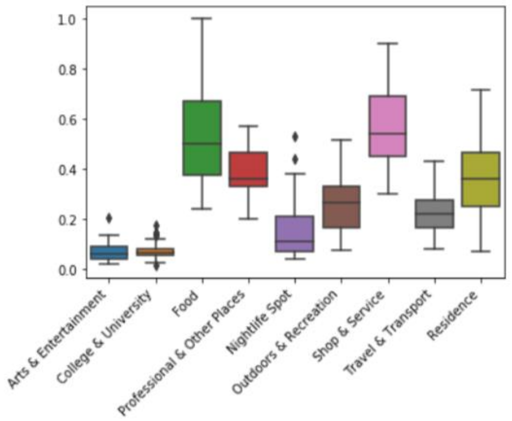
First, lets explore the map that occurs when two clusters are used:

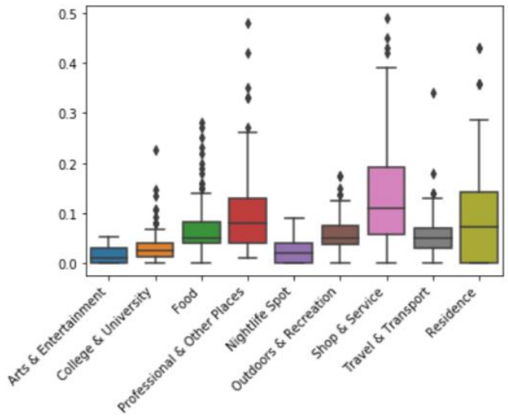


*Berlin divided into two clusters*

As expected, the city centre mainly forms the first cluster while the second one consists of the suburbs.

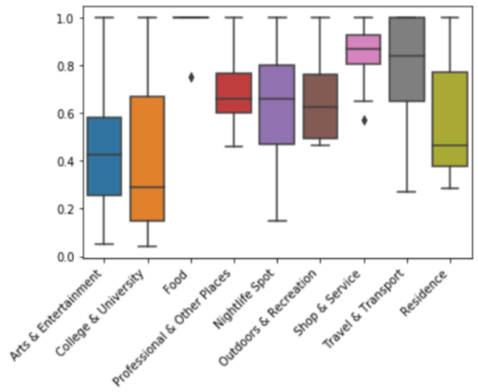
For the final analysis let’s settle on 3 clusters (0 to 2). Let’s visualize the clusters profiles using boxplots.





*Cluster 0 (red)*

*Cluster 1 (purple). Attention: differnet skale!*

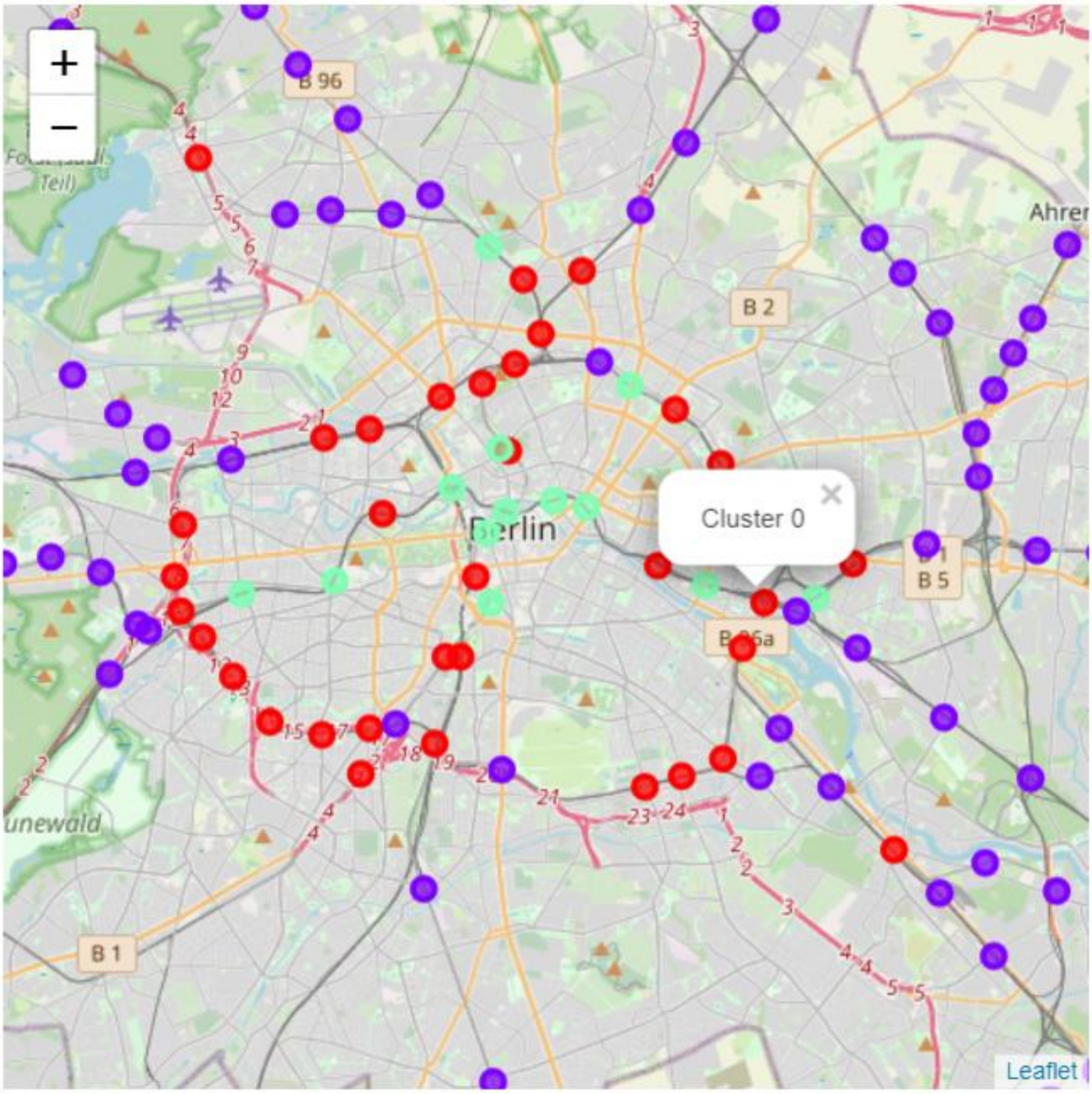


*Cluster 2 (green)*

Here is how we can characterize the clusters by looking at venue scores:

- Cluster 0 (red) has average marks at the most labels. While the Arts and Entertainment label and college and university are very low, food and shop and Service are high.
- Cluster 1 (purple) has low marks across the board. These appear to be underdeveloped areas.
- Cluster 2 (blue) has consistently high scores for all venue categories. This is the most diversely developed part of the city

The clusters are marked in the following map:



*Berlin divided into four clusters*

## Discussion

To be fair, Foursquare data isn't all-encompassing. The highest number of venues are in the Food and Shop & Service categories. Data doesn't take into account a venue's size (e.g. a university building attracts a lot more people that a hot dog stand — each of them is still one Foursquare "venue").

All in all, it was possible to divide the city into different clusters. This data can be useful for city planners to determine where from and where to people are most likely to travel for work and leisure, plan further extension of the network and find places for new development. As expected, there is a city centre with a consistently high scores for all venue categories. With same exemptions, the so called "ring" forms the next cluster. Almost everything is presented over there, but still in a smaller number than in the city centre. The last cluster consists of suburb areas that have low marks across the board. These appear to be underdeveloped areas.

## *Conclusion*

Foursquare data is limited but can provide insights into a city’s development. This data could be combined with other sources (e.g. city data on number of residents) to provide more accurate results.