IBM Data Science Capstone Project

Outline

- Target
- Data
- Methodology
- Clustering
- Result

Target

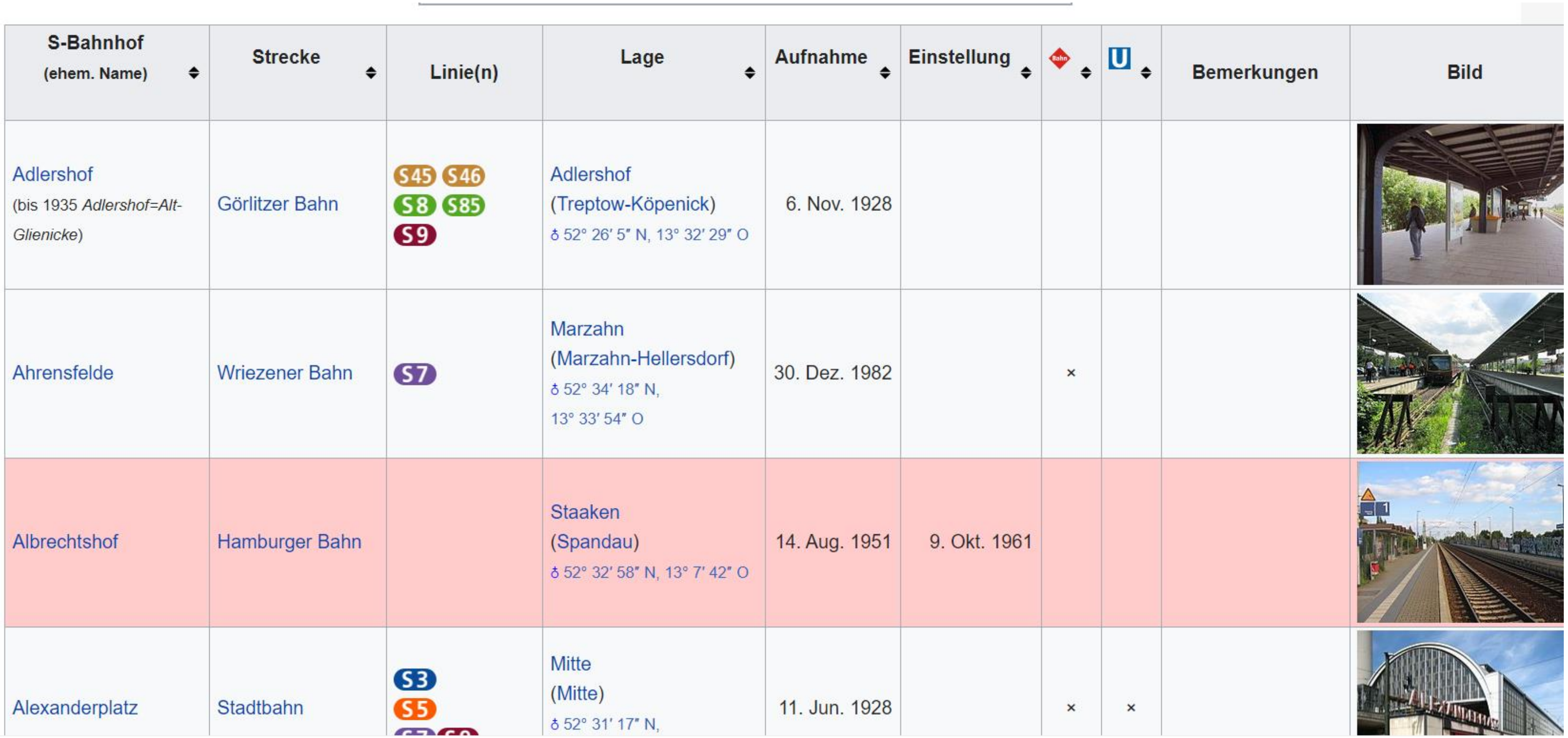
Group the stations of the S-Bahn of Berlin into different clusters.

The basis for this is the number of venues in the vicinity of the stations.

The venues will be divided into different categories.

Data

Geographic data of Berlins stations from Wikipedia 157 stations



S-Bahnhof (ehem. Name)	Strecke	Linie(n)	Lage	Aufnahme	Einstellung	Bahn	U	Bemerkungen	Bild
Adlershof (bis 1935 *Adlershof=Alt-Glienicke*)	Görlitzer Bahn	S45 S46 S8 S85 S9	Adlershof (Treptow-Köpenick) ♁ 52° 26′ 5″ N, 13° 32′ 29″ O	6. Nov. 1928					
Ahrensfelde	Wriezener Bahn	S7	Marzahn (Marzahn-Hellersdorf) ♁ 52° 34′ 18″ N, 13° 33′ 54″ O	30. Dez. 1982		×			
Albrechtshof	Hamburger Bahn		Staaken (Spandau) ♁ 52° 32′ 58″ N, 13° 7′ 42″ O	14. Aug. 1951	9. Okt. 1961				
Alexanderplatz	Stadtbahn	S3 S5	Mitte (Mitte) ♁ 52° 31′ 17″ N,	11. Jun. 1928		×	×		

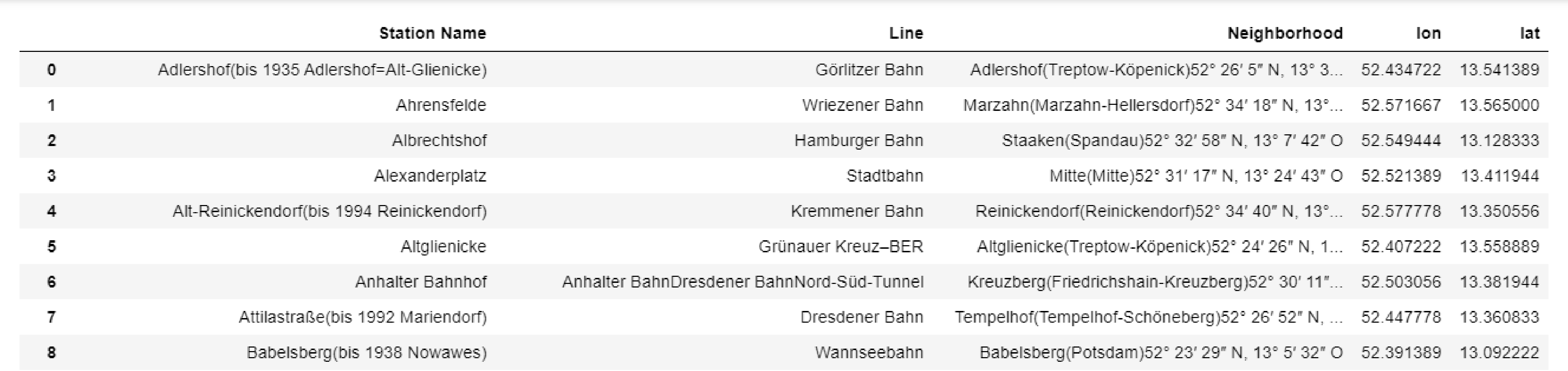
Data

Following steps were done to get a dataframe and clean the data

- Loud the table into Python with pandas
- 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
- Separating the minutes and seconds of the coordinates to transform them to a format that is supported by the 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 either in the longitude or in the latitude

Data

Cut-Out of the final datafram:



	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

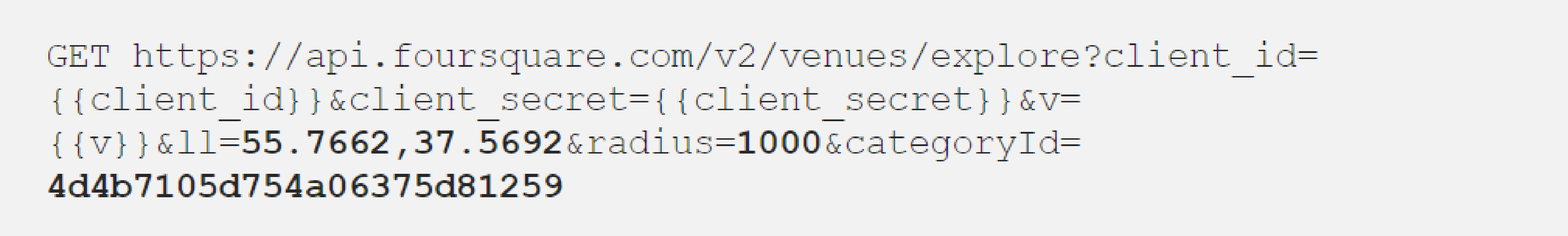
Data

Foursquare API to explore venue types surrounding each station. Foursquare outlines these high-level venue categories with more sub-categories. These are the feature for the clustering.

- 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)

Methodology

- We’ll be querying the number of venues in each category in a 1000m radius around each station.
- In Total 21350 venues
- Sample request (1000m radius and category Professional & Other Places):



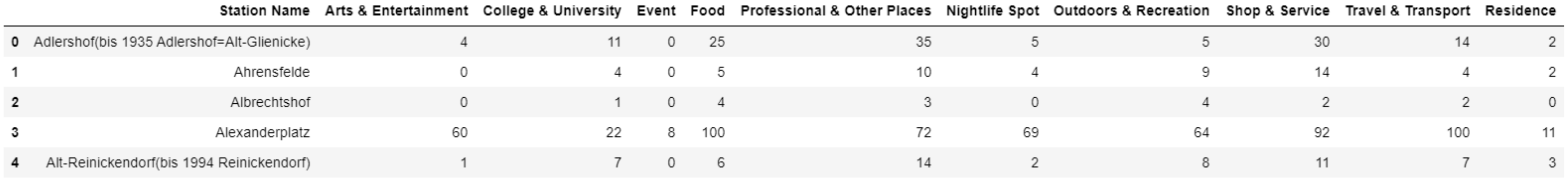
```
GET https://api.foursquare.com/v2/venues/explore?client_id=
{{client_id}}&client_secret={{client_secret}}&v=
{{v}}&ll=55.7662,37.5692&radius=1000&categoryId=
4d4b7105d754a06375d81259
```

Methodology

- Request all venues within 1000m for every station for every category
- Convert categorical variable “Category” with the ten categories into dummy/indicator variables
- Group the datafram by the station name
- Sum the numbers of every category for every station

Methodology

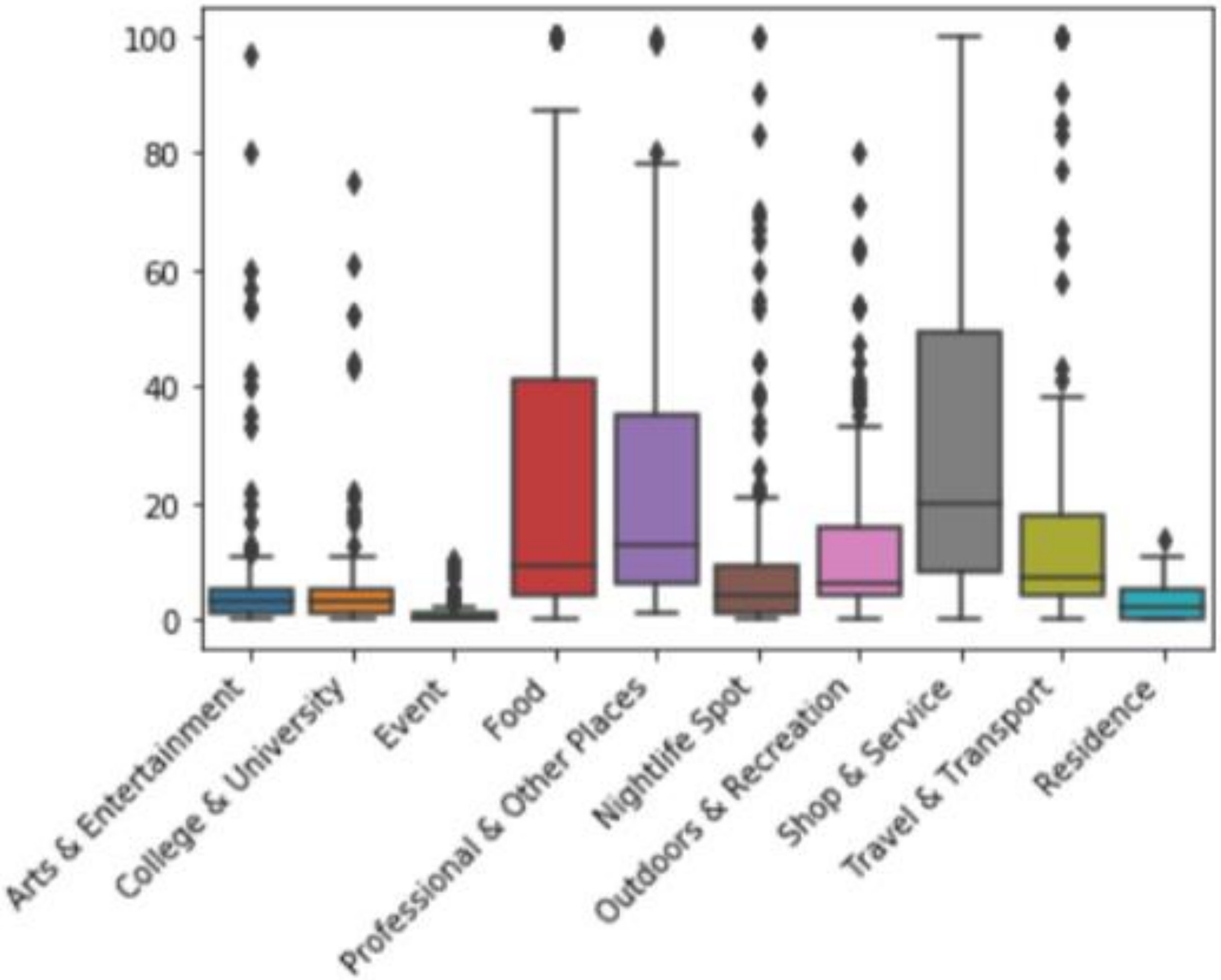
Cut-Out of the final datafram:



	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

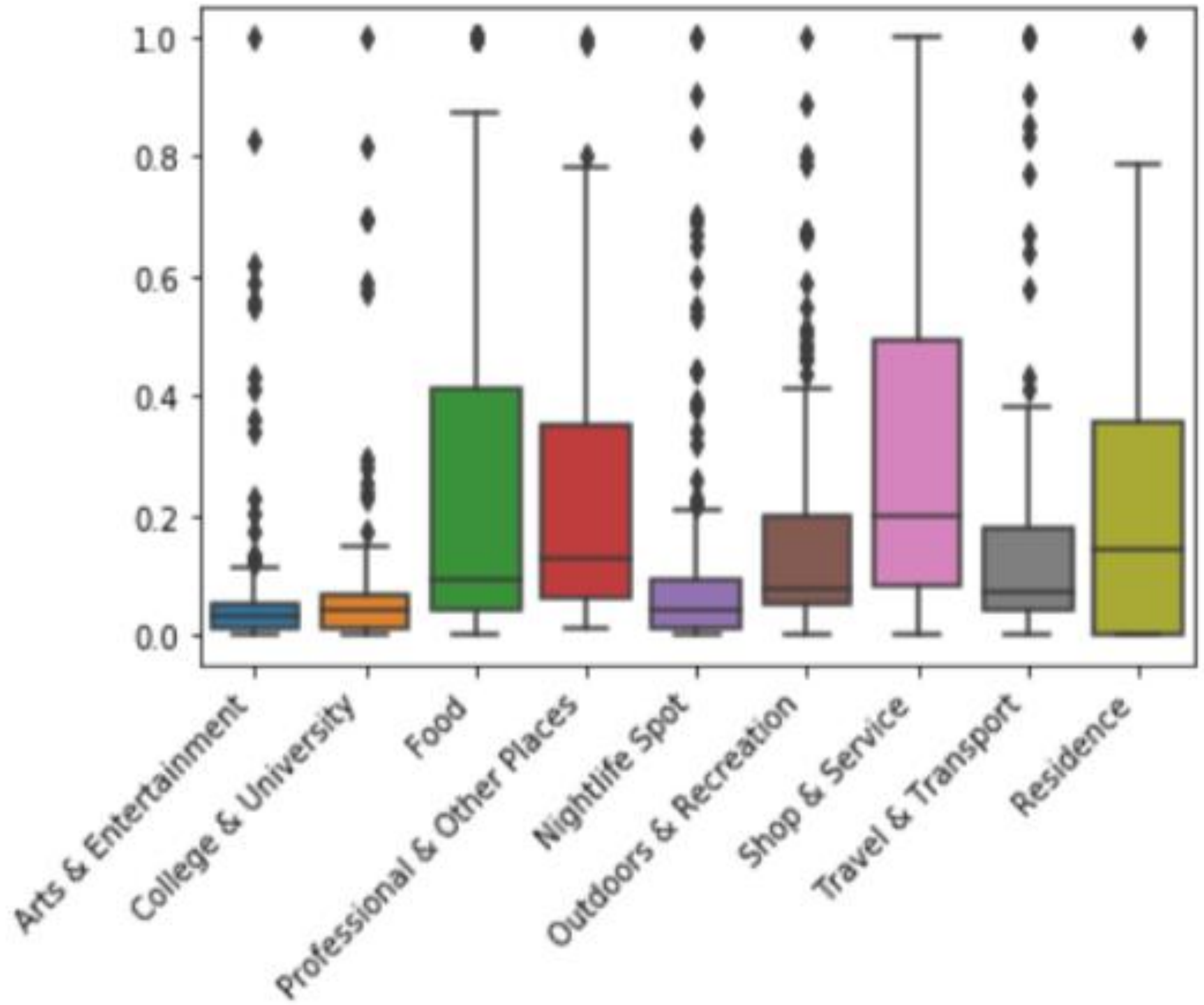
Methodology

Number of venues nearby the stations as boxplots (showing the average count, spread and outliers)



Methodology

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)

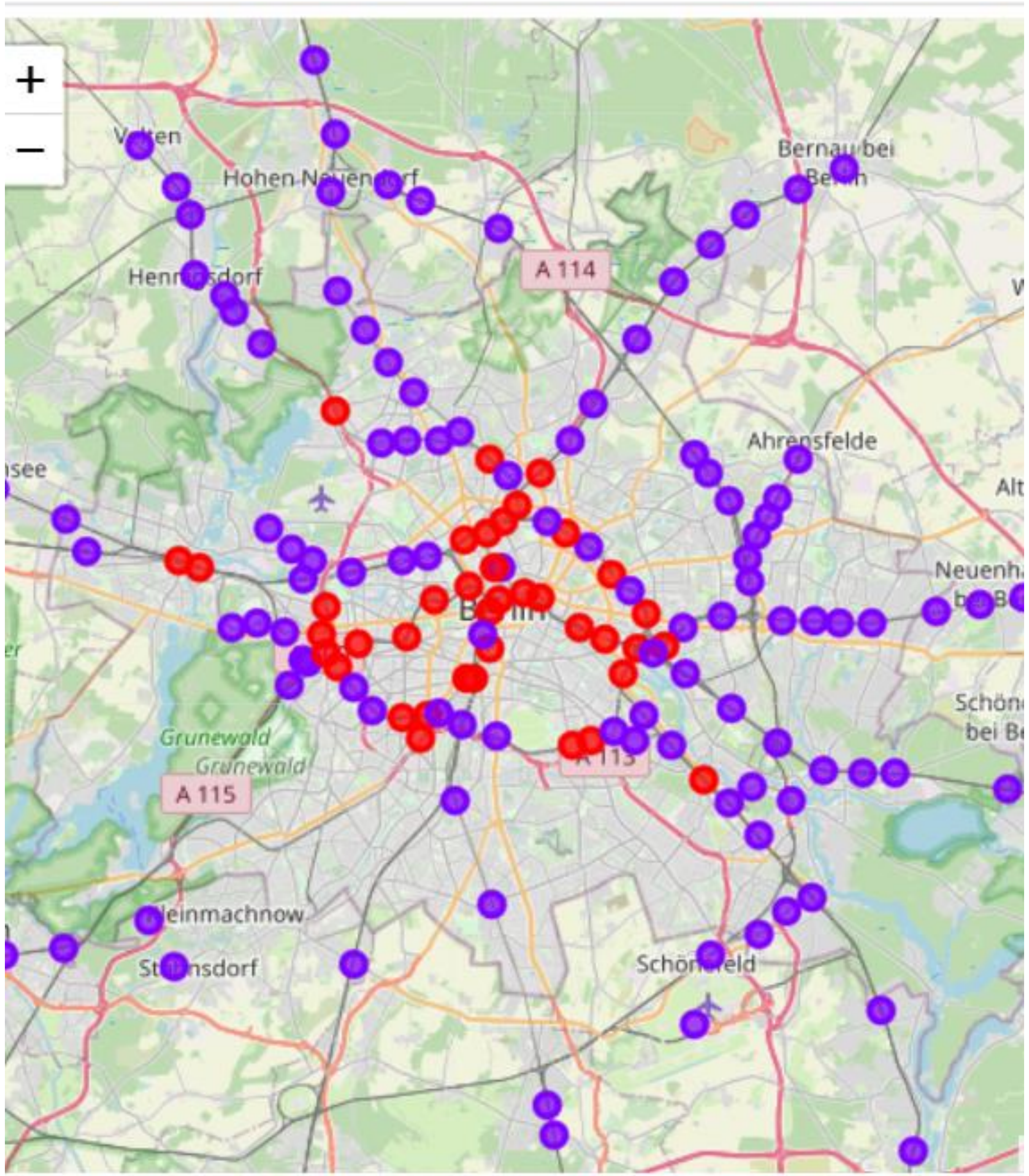


Methodology
Clustering

We’ll be using k-means clustering to cluster the stations by the number of venues in the surrounding area. The features are the normalized ten categories. These were the preliminary results with different number of clusters:

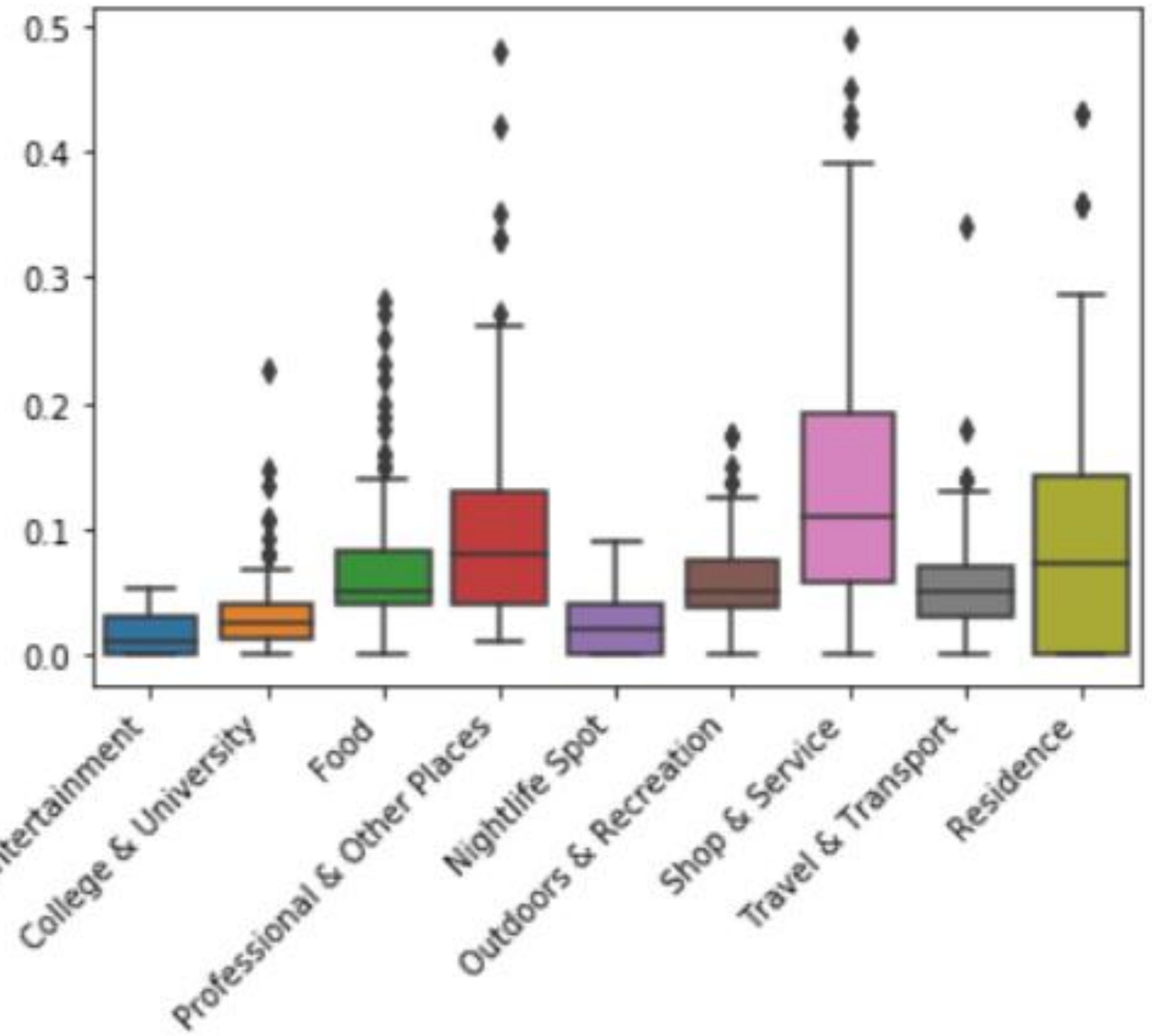
- 2 clusters: show the uptown/downtown divide
- 3 clusters: clustering within the downtown, also identify neighborhoods with very low number of venues

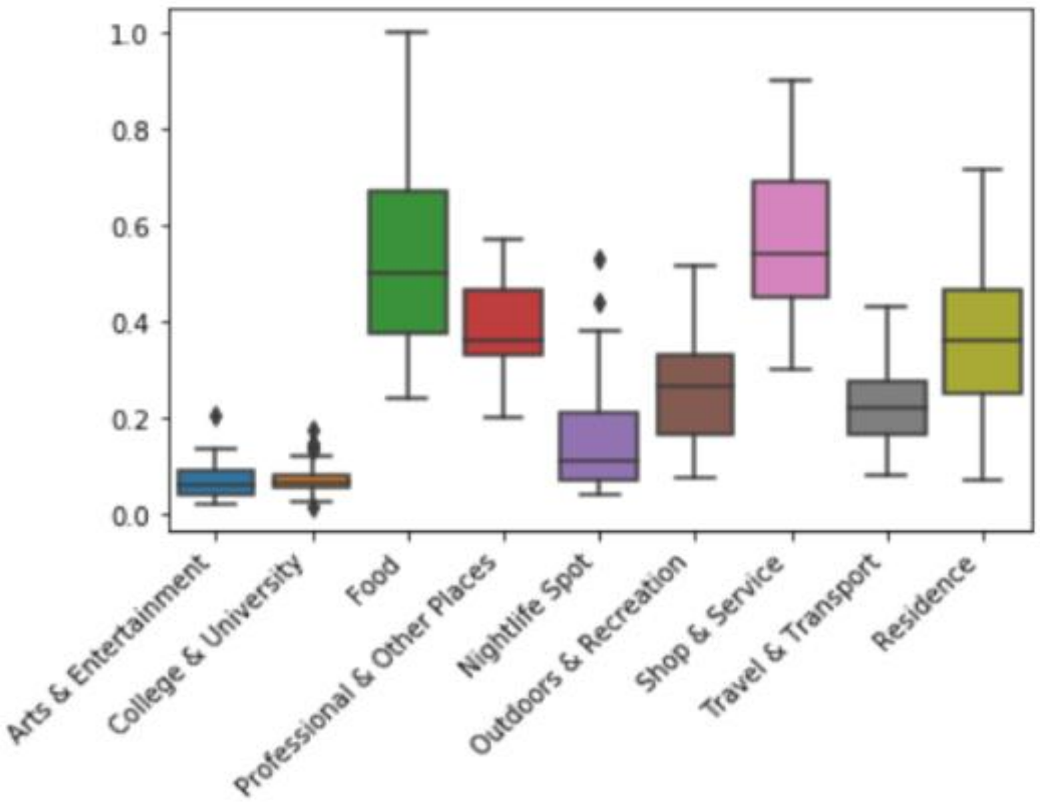
Result- Two clusters

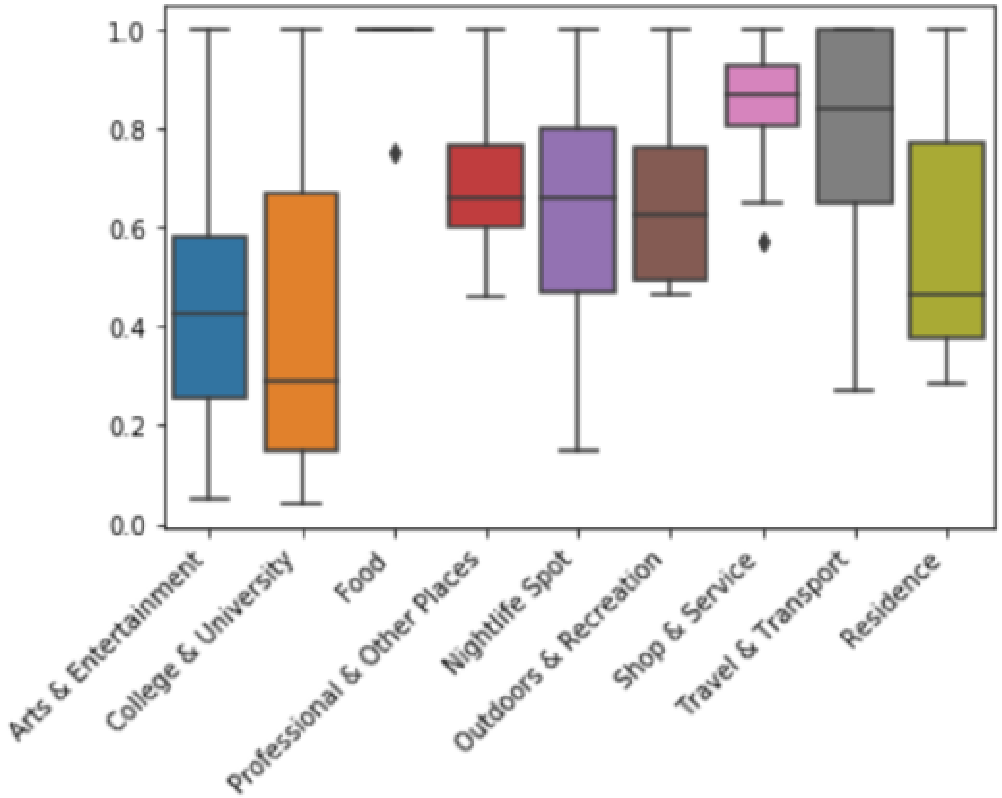


Result- Three Clusters

Cluster 1 (purple) has low marks across the board. These appear to be underdeveloped areas.





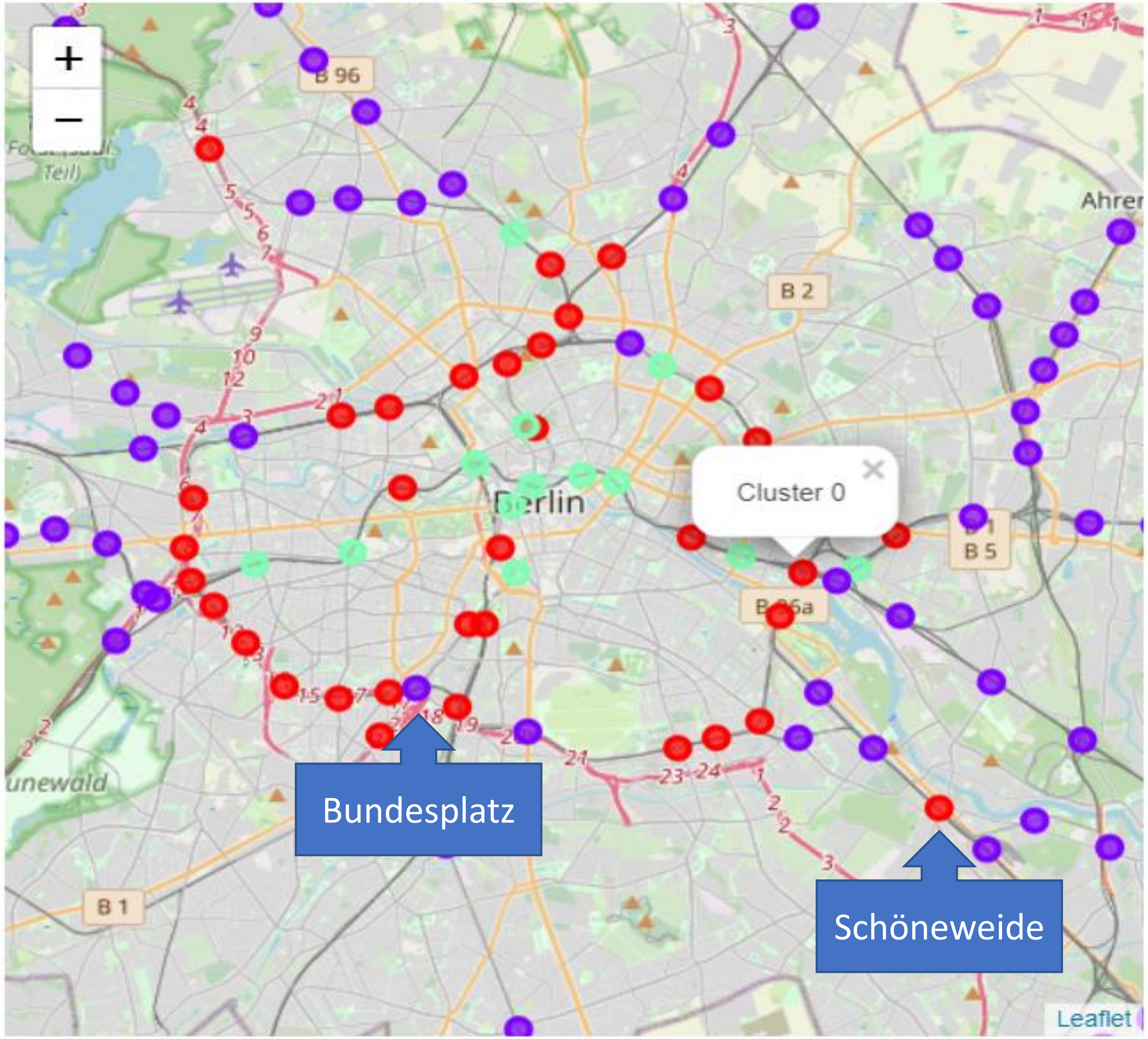


Cluster 0 (red) has average marks at the most labels. While the Arts and Entertainment label and college and university are very low, food and shop and Service are high.

Cluster 2 (blue) has consistently high scores for all venue categories. This is the most diversely developed part of the city

Result- Three clusters

- Cluster 0 (red)- underdeveloped areas.
- Cluster 1 (purple)- average marks at the most labels
- Cluster 2 (green)- city centre



Bundesplatz

Schöneweide