Project report:

Clustering Berlin metro stations using k-means algorithm and Foursquare API

by Tom Fenske
09.08.2020

Table of Contents

Table of Figures

1. Introduction

Metro lines in big cities are frequently used for commuting purposes or by traveling tourists who are sightseeing. This is relevant especially for cities like Berlin, which has 3.7 million inhabitants and in addition to that a remarkably high number of tourists. In 2019, Berlin was visited by approximately 13.7 million guests who spent 34.1 million guest-nights. The task for city officials is to run and maintain the transportation service in a way that there are no congestions in the transport capacity but also keeping the system cost-efficient.

Intuitively one might think that commuting traffic is highest on workdays, especially on mornings and evenings because that are the times when most workers travel to their working place or home from there. Also, public transport numbers for tourist purposes may be higher during the weekend than on weekdays due to day-travelers or short weekend vacations. So, it would be interesting to know where in the city the transport it used for what main purposes. Therefore, we might can cluster the stations of the public transportation net to learn more about where to expect what kind of traffic and draw conclusions on the expected load of the stations or lines.

Another use of this analysis might be found in the pricing/ticket system for public transport. If a cluster is found in the stations which is geographically compact and for example is interpreted as a mainly touristic cluster, city marketing might think of a special tourist ticket for this part of the city. With these tickets tourists maybe could travel within the designated cluster for a whole weekend (short trip) instead of buying all-day tickets for the whole city for every day of their stay.

Main problem of this analysis is to cluster the stations of a public transportation net to gain insights on the characteristics of the usage of the transportation net.

2. Data description

The data needed for this project would be a list of all Berlin metro stations. To perform a cluster analysis on this data additional information about the characteristics of each station is necessary. These characteristics might be the types of surrounding venues in the neighborhood of the metro station which are representing the possible destinations for travelers arriving at this station. For example, a metro station which has a lot of office buildings and cafes in its area could be interpreted as business category where a lot of commuting traffic is to be expected.

For this, a specific radius around each station must be defined in the analysis, in which the venues are lying. To get a list of venues around a specific location, Foursquare API can be used. Foursquare is a company which provides location-based data to allow users to explore their vicinity. It provides an API, whose 'explore'-function can be used for this purpose. This function needs to get passed the API credentials of the user, the area where to search (given by coordinates and a radius) and a maximum limit of venues to return. The response of the API call is a .json file with the information about the venues like name, category, and exact location.

In conclusion, the dataset needed to perform a cluster analysis would be a list of all metro stations in Berlin with their respective geographical coordinates and the venues lying in the defined area around the station. The list of metro stations can be retrieved from the corresponding Wikipedia page. The coordinates are found on this page, too, but they needed to be cleaned and modified to be passed to the API function. The venues for each station are responded by the Foursquare API call.

3. Methodology section

Starting point for the analysis is a Pandas dataframe which consists of the names of the Berlin metro stations and their corresponding latitudes and longitudes. With these coordinates, the Foursquare API 'explore'-function can be called. The response of the API call is a .json file with the information about the venues. The .json file then gets searched for the relevant info of the found venues.

These are written as a dataframe and merged with the station dataframe on the station name so that there is now a dataframe which consists of a list of station names and the venues found around this station. This dataframe is set up with dummy variables for the categories of the venues and grouped by station name so the resulting dataframe has one row for each station and the categories of the found venues in a radius around it represented by dummy variables. With this feature set, the k-means algorithm can be applied.

The k-means algorithm is a method of unsupervised learning where a feature set get categorized into different clusters whereas the number of clusters (k) must be predefined. In this analysis, the number of clusters is defined, and the k-means is responding a cluster label for each row of the feature set (which represents a single station). A new dataframe is created consisting of station name, coordinates, cluster label and a sorted list of the 10 most mentioned category values for each station for interpretation purposes.

It is inherent for k-means algorithm that the interpretation value of the analysis varies with strongly with the chosen k. As the user must define the value of k, a metric is needed to find out if the chosen value for k is useful. In this case the elbow method is applied. It runs the analysis with different values of k and measures the distortion value. The distortion value is the sum of squared distances from each data point to its designated cluster centroid.

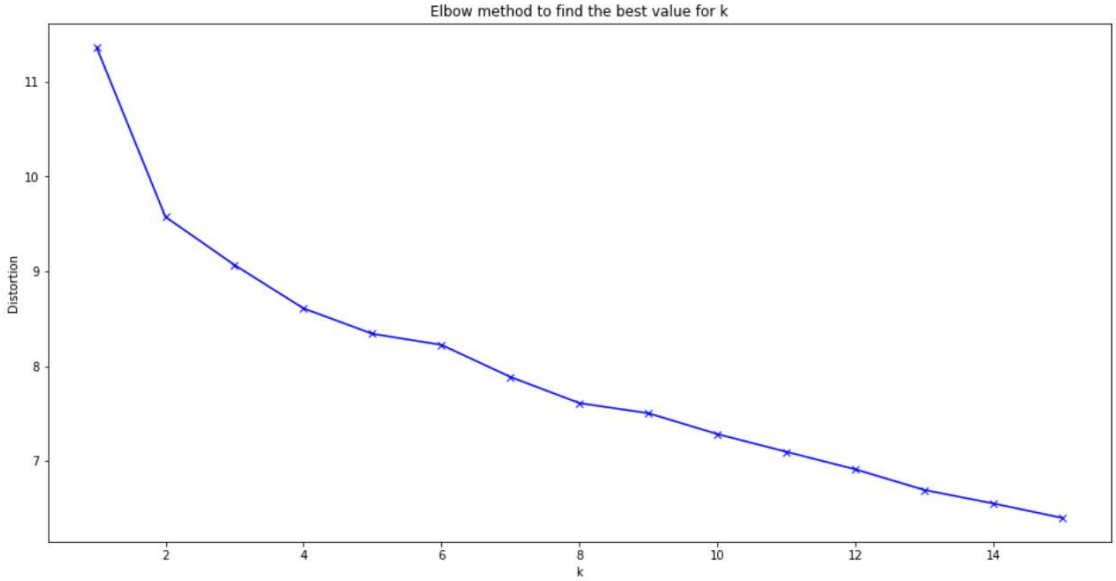

Figure 1: Elbow method to find the best value for k [Source: own graphic]

Usually the best value of k is found where the slope of the function decreases significantly (representing an 'elbow'). In this analysis, a sharp elbow is not found. Considering this, the complete analysis is done for different values of k and displayed accordingly.

To visualize the results, a folium map for each value of k is created. It is a map of Berlin, where every station is marked as a colored point. The color of each point is defined by the cluster label it is given.

4. Results section

The result of the analysis is a dataframe consisting of station name, coordinates, cluster label and a list of the 10 most found venues around this station. To visualize the results, a Folium map is created. The map shows the city of Berlin, whereas every station is displayed as a colored point at the respective coordinates. The color of the point depends on the value of the cluster label.

By this, areas of the city with stations are of similar characteristics (same cluster label) are particularly good to retrieve at first sight. Since the elbow method has not defined a clear result on what to set the value of k (see methodology section), the analysis is run for k equals 2, 5 and 8.

The first value of k is set to '2' which leads to the following Folium map:

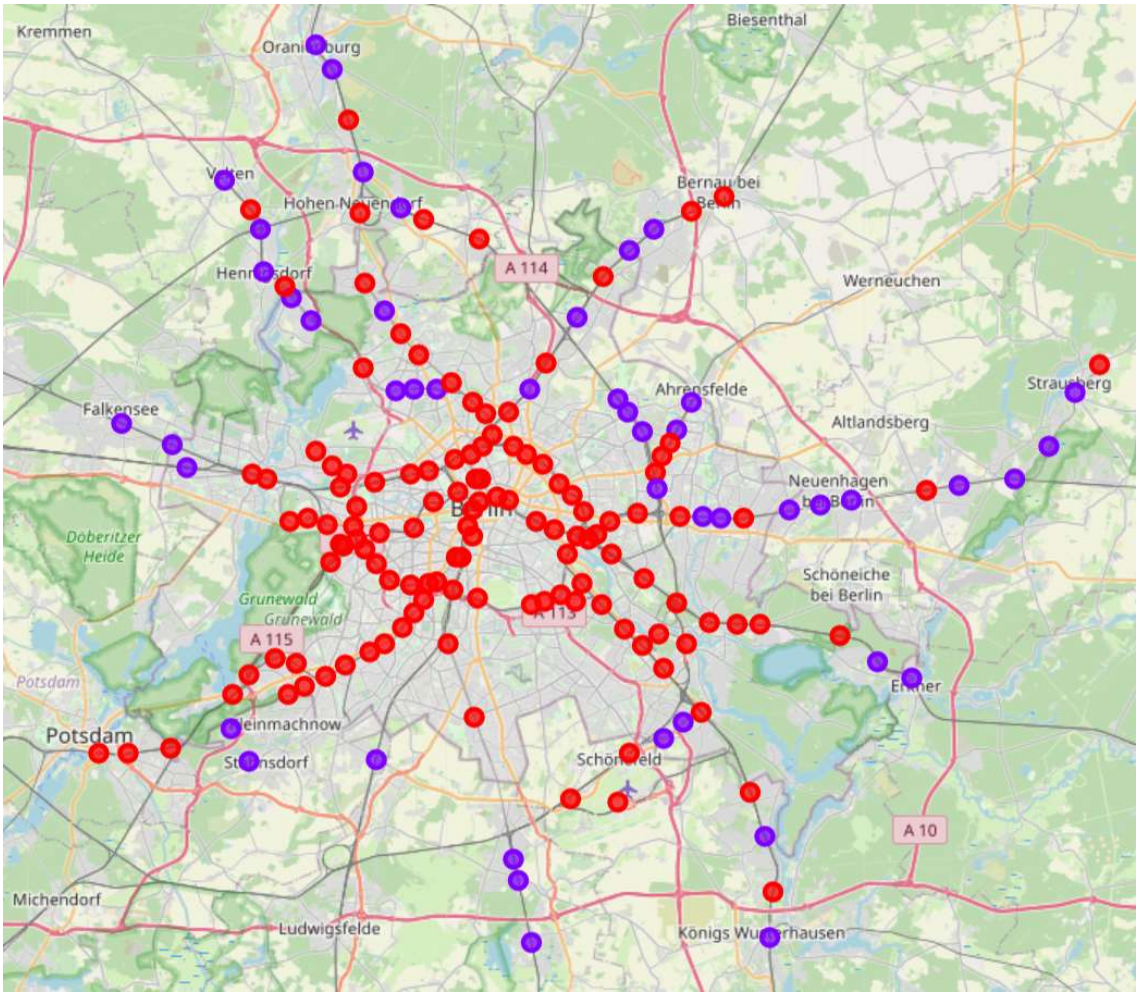

Figure 2: Folium map of Berlin metro stations using 2 clusters [Source: own graphic]

Red points are labeled as cluster 0 and purple points are labeled as cluster 1. The interpretation of the clusters by the surrounding venues implies a commercial/tourist type of area for the red points and a residential type for the purple points. It is plausible at first sight that people are living in the outside areas of the side or even out of the city borders whereas the inner parts of the city are characterized by business and tourist venues.

The next run was done with a k value of '5' which lead to the following Folium map:

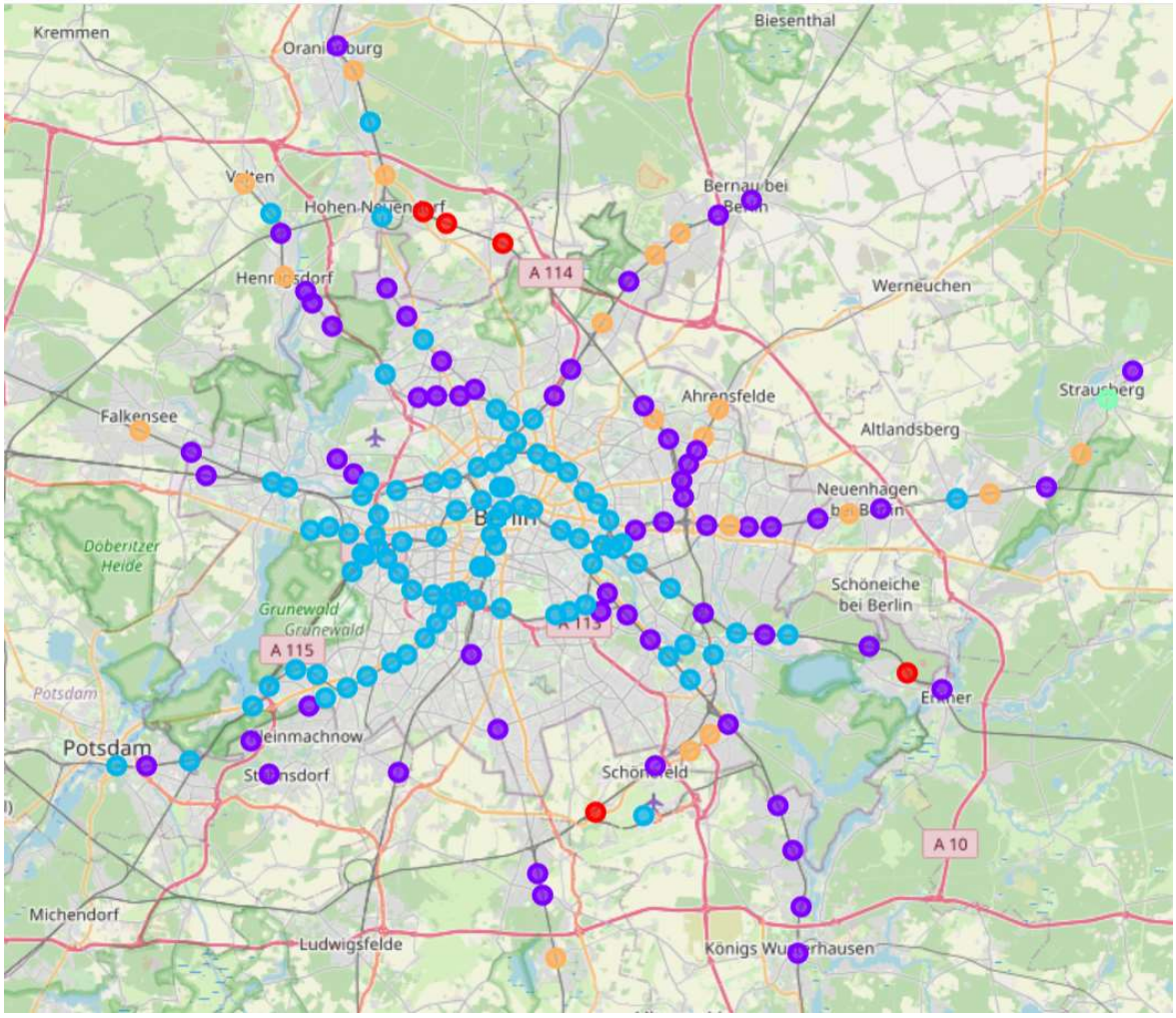

Figure 3: Folium map of Berlin metro stations using 5 clusters [Source: own graphic]

Cluster 0 is represented by the purple points and is interpreted as a residential type of area. Cluster 1, which is colored turquoise, is interpreted as a commercial respectively touristic area. Cluster 2 is dismissed from the analysis because it contains only one station. Cluster 3 is also a residential type and is shown as orange points, whereas Cluster 4 is considered a rural type of area and marked as red points.

The last run was done with a k value of '8' which lead to the following Folium map:

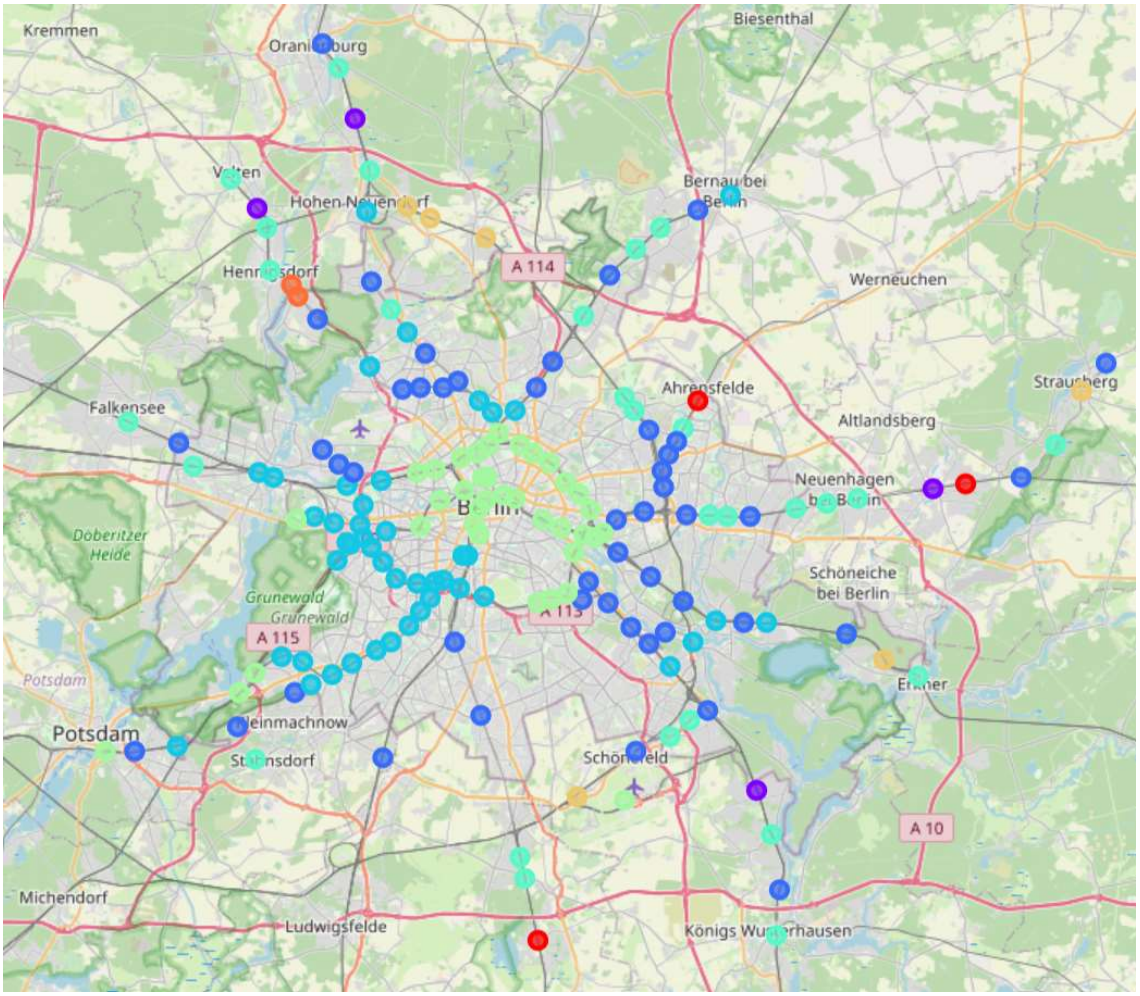

Figure 4: Folium map of Berlin metro stations using 8 clusters [Source: own graphic]

Cluster 0 is represented by purple points and is interpreted as rural type. Residential areas are found at the blue points (cluster 1) and on the turquoise points (cluster 3). Cluster 2 (cyan points) has also some residential characteristics, but with a commercial influence, whereas cluster 4 (light green points) is a tourist type of area. Yellow points are labeled as cluster 5 and are interpreted as industrial type of area. Cluster 6 (orange points) and Cluster 7 (red points) are dismissed because they only contain 2 (respectively 3) stations.

5. Discussion

The results of the analysis in form of Folium maps are showing the distribution of the clusters over the city depending on the specific characteristics of the area around each station. The distribution supports the intuition of residential areas lying near the city borders or outside of it, whereas commercial and tourist areas are mainly distributed inside the metro ring. This can be seen especially when choosing a little value for k, here '2'. Choosing higher values of k refines the distribution of cluster labels inside the metro ring. Now even touristic and commercial areas can be separated from each other. Also, on the outside of the city residential and rural areas can be separated.

For further analysis it can be considered the run the algorithm on smaller parts of the city like a borough or even a neighborhood. With this, maybe other types of public transport lines like tram or bus stations can be added to gain deeper knowledge of the characteristics of public transport in a defined area of the city.

Finally, a different cluster algorithm like DBSCAN could be used for to re-run the whole analysis and then compare it with the results of k-means algorithm or find some additional information.

6. Conclusion

The goal of the analysis was to learn about the different characteristics of Berlin metro stations to gain knowledge about expected load for city officials and/or insights to use in the pricing system for city marketing.

The analysis has shown that stations outside the metro ring and outside the city are mainly residential areas so the traffic can be interpreted as commuting traffic. With this, the load can be expected to be highest on workdays, especially in the morning and the evening.

The stations inside the metro ring are mainly interpreted mainly as commercial and/or tourist types. Therefor is could be considered to define a special ticket for inside the metro ring that is more cost-efficient than a whole-day ticket to make it attractive for tourists.