The Battle of Neighborhoods: Pandemic Friendly Metro Stations in Berlin

By Umrbek Allakulov

Introduction and problem statement

Pandemics, such as COVID-19, tend to spread more easily in cities as the transmission of infections is facilitated by population density and intensive use of public spaces. In addition, global cities and tourism hotspots also facilitated the spread of the novel coronavirus, as people tended to travel to and from these destinations before the patients could show any symptoms of the infection.

For people moving to global cities such as Berlin during the unprecedented times of the coronavirus pandemic, it can be a daunting task to decide in which neighborhoods to look for housing. That means, on the one hand, newcomers will likely prefer to move to neighborhoods where they have access to basic goods and services within walking distance, as that would reduce the need for using public transport and diminish the risk of contagion. In addition, they may also wish to reduce the likelihood of coming into contact with people who recently travelled from abroad by looking for housing in areas where fewer tourists visit.

In this project, which I am undertaking as part of my Applied Data Science Capstone, I attempt to answer the following question:

Which metro stations in Berlin are located in neighborhoods that are more convenient and safer for living in during episodes of pandemics?

Data

To answer the above question, I will require venue and geolocation data which can be used to cluster the neighborhoods hosting the metro stations into similar and dissimilar groups. The following are the datasets I have identified for the purpose of this project:

1. The Foursquare Places API data

The dataset contains data on 6840 venues. The variables of interest are the venue name, categories, and geolocation. One limitation of this dataset is the limited number of observations per station. The Foursquare API has imposed a limit on the number of regular calls, which I set at 30 per metro station.

2. Coordinates of metro stations belonging to the ***Berliner Verkehrsbetriebe*** ("Berlin Transport Company")

I obtained the dataset on the metro stations in Berlin on the official site of the Berlin city administration: https://www.vbb.de/media/download/2035
Overall, the dataset contains information on 293 metro stations including the underground and elevated stations (i.e. U-Bahn and S-Bahn).

Methodology

To identify neighborhoods where it may be more convenient and less risky to live in during episodes of pandemic, I have selected 6 categories. The following 3 categories represent venues that offer essential goods and services, which make neighborhoods more convenient during lockdowns:

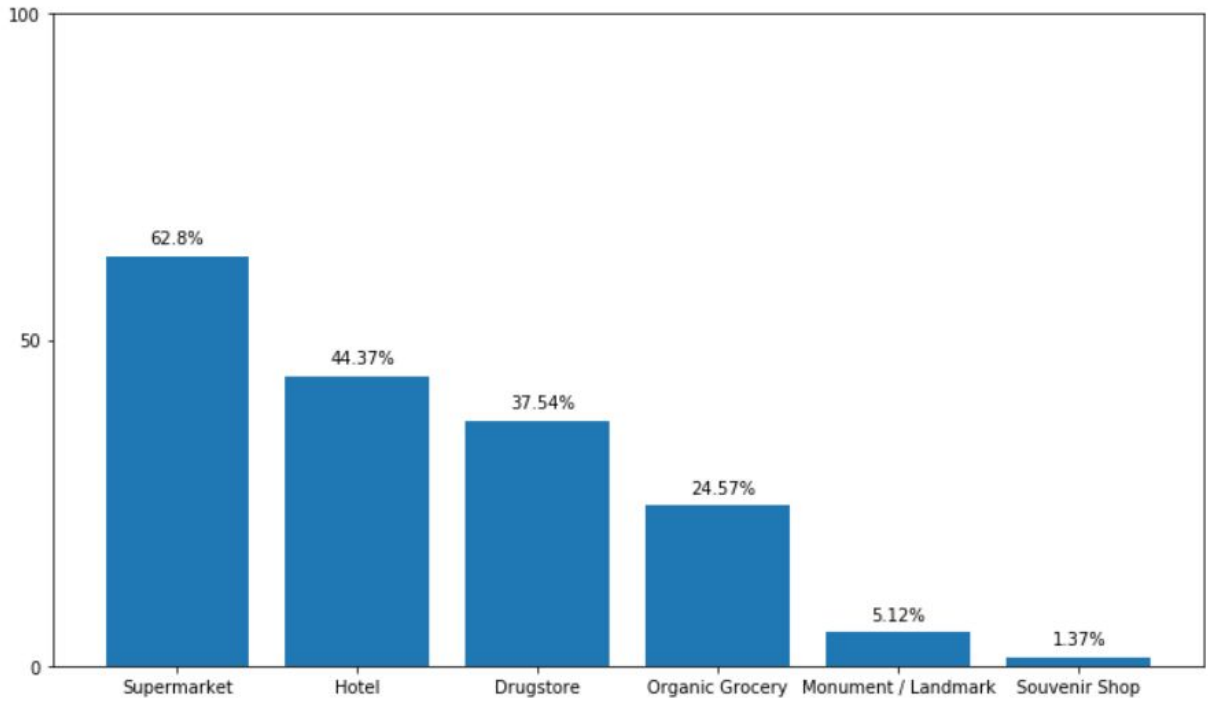
1. Supermarket
2. Organic Grocery
3. Drugstore

Additional 3 categories represent venues that offer non-essential good and services for locals and appeal to tourists, thus making the neighborhood more risky:

1. Souvenir Shop
2. Monument / Landmark
3. Hotel

Descriptive statistics

As can be seen in Graph 1, almost 63% of stations have at least one supermarket within 500 meter radius. Less than half of stations (44%) have at least one hotel nearby. On the other hand, very few stations have at least one monument or souvenir shop nearby (approximately 5% and 1% of stations, respectively).



Graph 1. Proportion of stations containing at least one of selected venue categories

Clustering the metro stations

To divide the neighborhoods into similar and dissimilar sets, I have used k-means clustering method, which is an unsupervised machine learning technique used to identify clusters of data objects in a dataset. Setting the partitioning parameter k to 3 gives us three distinct clusters of neighborhoods with the following characteristics:

Table 1 shows the results of the partition.

Cluster Labels	Drugstore	Hotel	Monument / Landmark	Organic Grocery	Souvenir Shop	Supermarket
0	0.000000	0.333333	0.000000	0.092593	0.009259	1.518519
1	0.311927	1.963303	0.155963	0.403670	0.027523	0.073394
2	1.355263	0.355263	0.013158	0.328947	0.000000	1.750000

Table 1. Three clusters of neighborhoods

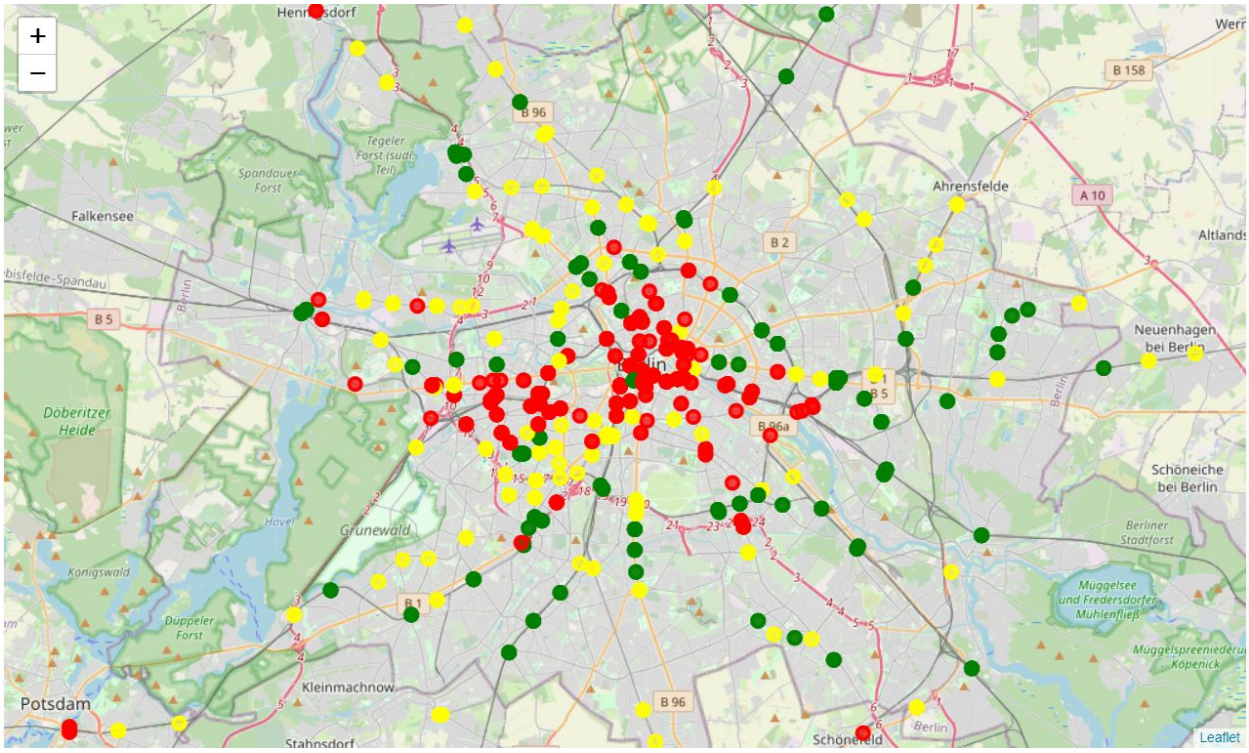
- Cluster 0
    - Contains metro stations with relatively high concentration of supermarkets and relatively low concentration of hotels, organic grocery stores, and souvenir shops.

There are no drugstores or monuments (landmarks) in the neighborhoods belonging to this cluster.

- Cluster 1
    - Contains the highest concentration of hotels, monuments, and souvenir shops, suggesting that these are popular neighborhoods for visitors. This cluster also has the most number of organic groceries. However, Cluster 1 neighborhoods enjoy relatively low concentration of supermarkets and drugstores
- Cluster 2
    - Features neighborhoods with the highest concentration of supermarkets and drugstores. These neighborhoods also often have an organic store in proximity. In addition, Cluster 2 features relatively few hotels and landmarks, and no souvenir shops. It is easy to conclude from these characteristics that Cluster 2 neighborhoods are residential parts of the city.

Discussion of results

As can be seen in the map (here is a link to the interactive map), most green areas are located at the outskirts of the city. The risky (red) neighborhoods are concentrated in central parts of Berlin, which are popular among tourists. The yellow cluster is an interesting one, as it combines some locational shortcomings as well as benefits. More specifically, there is less concentration of essential venue types around the yellow neighborhoods compared to green ones. However, they are also less appealing to tourists than red areas and benefit from proximity to the central parts of Berlin.



Graph 2. Map of neighborhoods in Berlin

It is also worth noting that the red cluster picks up relatively more spots stretching in the direction of west Berlin, which is historically a more affluent part of the city and now hosts more tourist attractions.

Conclusion

The presented analysis should be used with due care, as data limitation emanating from the Foursquare API call limit may have caused possible errors in clustering. However, the identified clusters can serve as an initial shortlist of neighborhoods for those searching for more convenient and less risky areas of Berlin to move to during the ongoing pandemic.

The interactive map with the clusters is available here.