Berlin venues

and the spread of the coronavirus

Are there certain types of urban structures that increase the risk of infection?

Kristin Lehnert

December 2020

**Introduction**

The corona virus has kept the world on edge all the past year. Many are currently hoping that the recent approval of the first corona vaccine in the EU will bring a rapid improvement. However, at this moment is still unclear how quickly it can be delivered, who will be supplied first and how many Germans will actually be willing to be vaccinated in the end. The persistent uncertainty has made many people cautious[[1]](#footnote-1). In the meantime, the numbers continue to rise in Berlin and everywhere else.

The city of Berlin in particular has been the subject of much criticism from the German government in recent months[[2]](#footnote-2). The beginning of the Christmas season was doused in some districts on the streets with mulled wine. Even if stores and restaurants had closed, people met at their old usual places – especially places that are known to be busy from “normal” times and that are popular with young people for their rich leisure offer. The districts of Neukölln and Mitte were the first to see case numbers exceed the critical range[[3]](#footnote-3).

But is there a particular reason why the virus is spreading faster in certain districts than in others? Is there even such a thing as a “risky infrastructure”? And if so, are these just a few or various places in the city? In this investigation, Berlin districts were clustered based on their venue infrastructure and placed in the context of Covid case numbers. The aim was to analyze whether a certain constellation of places where people come together may have an influence on the infection figures in this region.

The problem was studied as part of the capstone project for the IBM Professional Certificate in Data Science.

**Data**

To approach the problem, the following data sources were consulted, pre-processed and analyzed.

Table 1 overview of data and sources

|  |  |  |
| --- | --- | --- |
| List of Berlin districts and boroughs | Wikipedia | The table was scraped from https://de.wikipedia.org/wiki/Verwaltungsgliederung\_Berlins using an html-parser for conversion into a pandas data frame |
| Geospatial data of all districts | geopy | This data was obtained from the free Open Streetmap-API https://geocoder.readthedocs.io/api.html#forward-geocoding |
| List of venues in each district | Foursquare | A json file with all district’s venues was taken from Foursquare, a proven independent location data platform https://developer.foursquare.com/ |
| Current covid infection figures | Berlin Open Data | The current Corona infection figures were retrieved from the official capital portal of the city of Berlin https://daten.berlin.de/datensaetze/covid-19-berlin-verteilung-den-bezirken |

**Methodology**

The first task was to determine the specific types of venues offered by the individual districts of Berlin and how the city can be clustered on the basis of these similarities between the various districts.

Starting point for the analysis was a pandas data frame of Berlin's 96 districts, divided into 12 boroughs. After scraping the data from Wikipedia, the dataset was preprocessed in Python, merged with its longitude and latitude coordinates using the geopy library and visualized in a Leaflet map via folium.

Ein Bild, das Karte enthält.

Automatisch generierte Beschreibung

Figure 1 Districts of Berlin

These coordinates of the districts were then used to call the Foursquare API 'explore' function, which returned a json file containing the location information. The file was searched for "venues", so that a dataframe with all Berlin districts and their venues was available as a result.



Figure 2 Excerpt from the list of districts, venues and geo data

238 individual categories were found, most of them in Neukölln (74 venue categories), followed by Moabit (64), Kreuzberg (55) and Mitte (54). Venues on Foursquare are all public meeting places, from train stations to fitness studios, from the supermarket to the small snack bar around the corner. In total, 238 individual categories of venues were found in Berlin.

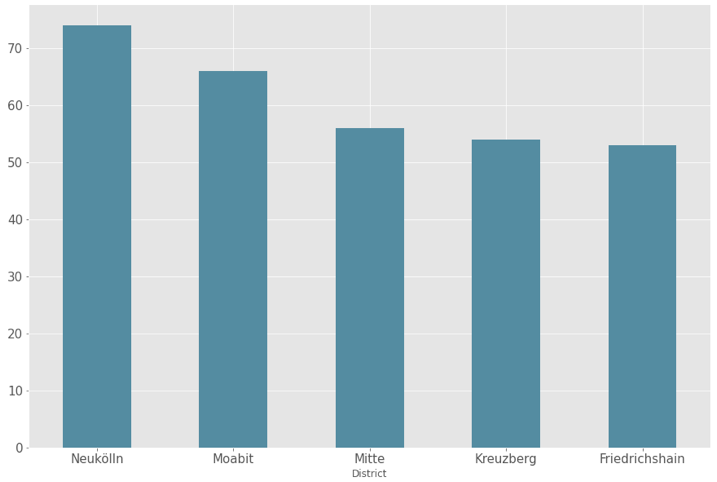


Figure 3 Berlin districts with the highest number of venues

From a descriptive perspective, comparing this graph to the Corona virus incidence in the boroughs a certain semblance is recognizable. At first glance, this might suggest that the number of venues in a district is related to the number of corona infections. This question will be hypothesized and statistically tested within the scope of this investigation.

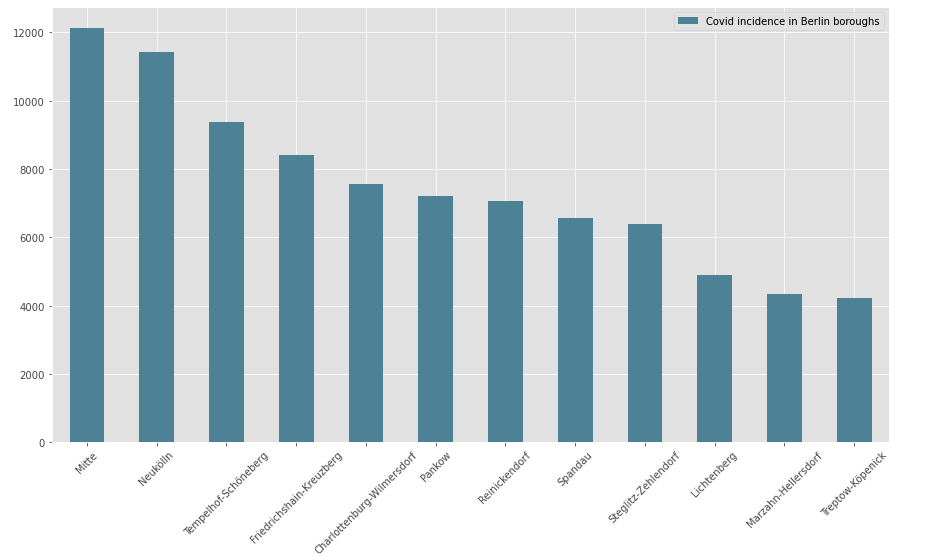


Figure 4 Berlin districst with the highest Corona incidence

The next step was to find out which are the most present venues in each district. For this purpose, all categories they were dummy-coded so that the resulting data frame contained one row for each district and the availability of every sort of venue there. Based on this, the top 10 venues for each district were identified.

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

Figure 5 Excerpt from the list of districts and their most common venues

After the data had been prepared to this point, finally an algorithm was applied to cluster the districts according to their characteristic venue-structure. This purpose was met by the k-means algorithm, a method of unsupervised learning that automatically categorizes a feature set into different clusters. Thus, the similarity of the neighborhoods in terms of their venue offer was determined by the algorithm, while the interpretation of the resulting groups was still up to the investigator. It is inherent in the k-means algorithm that the interpretive value of the analysis varies greatly with the k chosen. In this study, various values of k were tried and k=6 was selected as the most plausible clustering.

The resulting clusters were again merged with the location data and can now be identified by the different colors in the folium map.

Ein Bild, das Karte enthält.

Automatisch generierte Beschreibung

Figure 6 Berlin districts colored according to their assigned cluster

After splitting the dataset by venue clusters, certain specific qualities were identified for each cluster - according to free personal, noteworthy. The results of this closer look at the clusters are listed in the following table.

Table 2 Description of the 5 clusters identified by the algorithm

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cluster | Cluster Name | Example District | no. of districts | Example venues | Color |
| 0 | “the village-like cluster” | Pankow, Rosenthal | 13 | Automotive Shop, Miscellaneous Shop | red |
| 1 | “the vibrant nightlife cluster“ | Neukölln, Mitte | 59 | Café, Restaurant | violet |
| 2 | „the supermarket cluster“ | Mariendorf, Rudow | 14 | Supermarket, Park | blue |
| 3 | “the calm neighborhood cluster“ | Kaarow, Malchow | 4 | Bus Stop, Playground | cyan |
| 4 | „the ‘nothing going on’ cluster” | Plänterwald, Rahnsdorf | 3 | Office, Bus Stop | green |
| 5 | "the beautiful living cluster" | Heiligensee, Kladow | 2 | Flower shop, Zoo | orange |

Now that we have found out how many different venues there are in each district of Berlin and also clustered them according to their venue structure with the help of AI, the next interesting question is how these conditions relate to the incidence of covid infections. Two hypotheses were formulated for this purpose.

Table 3 Hypotheses for the following investigation

|  |
| --- |
| 1) There is a positive correlation between the total number of venues in a borough and the incidence of covid infections. |
| 2) There is a positive correlation between the number of "vibrant" clusters in a borough and the incidence of covid infections. |

Unfortunately, the publicly available case numbers are currently only provided at borough level, not for each district. Therefore, adjustments to the data set were again necessary. The number of venues was cumulated at borough level and merged with the current dataset on corona infections in Berlin.

Furthermore, it was analyzed how many "vibrant" neighborhoods there are in each borough. Cluster 1 was chosen as a reference, as it is the cluster with the richest supply of leisure activities and also the most widespread cluster. Here is an overview of how many Cluster 1 neighborhoods there are in each berlin borough:



Figure 7 Berlin districts ranked by the number of "lively" neighborhoods

To examine possible relations with Corona infections, a regression analysis was performed. However, the two predictors „total number of venues" and "number of cluster1 districts" could not be combined in one multiple linear regression as both are based on the venue data and multicollinearity has to be avoided. Therefore, both hypotheses were tested individually using simple linear regressions.

**Results**

The number of cluster 1 districts was found to have no significant effect on the corona incidence in a region (p = .38, alpha = .05). The number of venues in a region also did not have the expected effect: more venues in a borough does not necessarily result in more infections with the virus (p=.06, alpha = .05) even though a slight trend is evident in the data.

Thus, the results of this analysis do not suggest that the differences in the risk of contagion between boroughs can be related to their specific mix of venue offerings.

**Discussion**

Obviously, there are several limitations to this analysis. First, the size and population density of the districts under consideration were not taken into account, which are certainly decisive factors. Second, with the cumulative numbers, there were not that many data points available to feed into the regression analysis, resulting in a limited validity of the results. In addition, most venues had been locked-down for months anyway, making it hard to detect their influence. Most infections currently occur at private gatherings, the frequency of which is not recorded in any statistics.

Nevertheless, this exercise has hopefully demonstrated an interesting approach to the topic and could possibly be applied in a similar way to other cities. The project mainly served to consolidate my new data science skills acquired through the IBM Data Science course and I certainly had fun doing it. The notebook with the complete evaluation is available at this [link](https://nbviewer.jupyter.org/github/Kristinlehnert/Coursera_Capstone/blob/main/Assignment_4/Berlin%20venues%20and%20the%20spread%20of%20the%20coronavirus.ipynb).

Link to Jupyter Notebook:

<https://nbviewer.jupyter.org/github/Kristinlehnert/Coursera_Capstone/blob/main/Assignment_4/Berlin%20venues%20and%20the%20spread%20of%20the%20coronavirus.ipynb>

1. https://www.aerzteblatt.de/nachrichten/118522/Laut-Umfrage-wollen-weniger-Menschen-definitiv-Impfung-gegen-Corona [↑](#footnote-ref-1)
2. https://www.welt.de/vermischtes/article217032028/Berlin-voller-Corona-Hotspots-was-ist-nur-los-in-der-Hauptstadt.html [↑](#footnote-ref-2)
3. https://www.tagesspiegel.de/berlin/deutsche-corona-hotspots-vier-berliner-bezirke-gelten-inzwischen-als-risikogebiete/26239218.html [↑](#footnote-ref-3)