Berlin venues

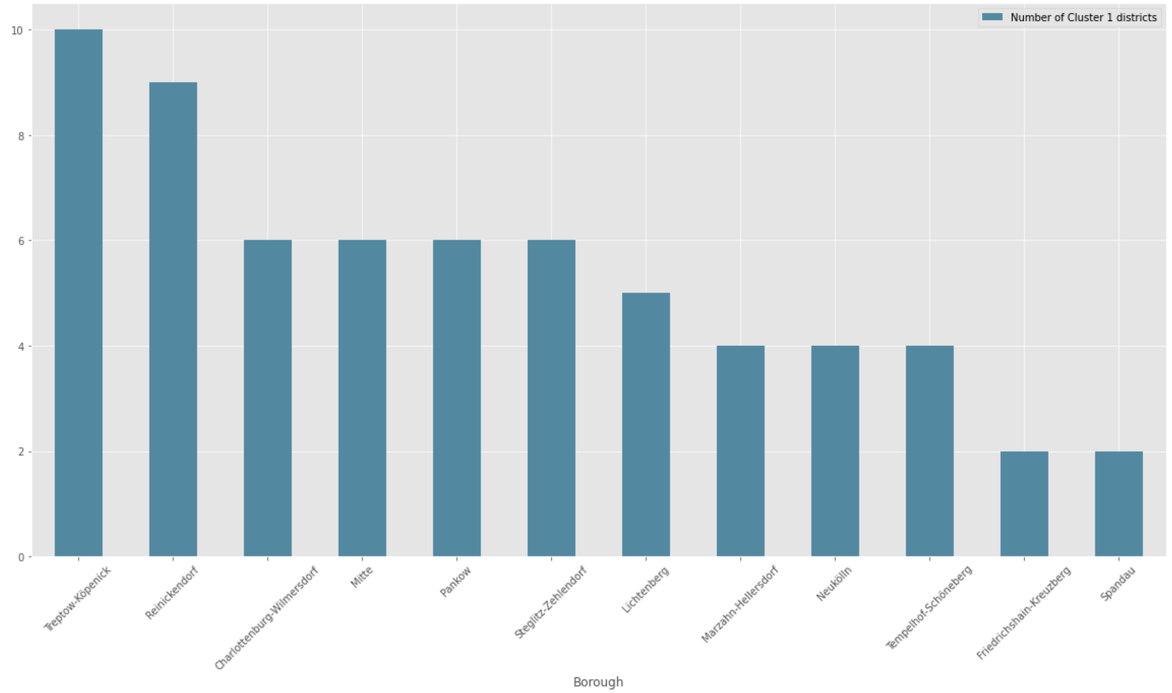
and the spread of the corona virus

Kristin Lehnert, December 2020

Introduction

Current Situation in Berlin

- Corona infections continue to rise everywhere, so in Berlin
- Some districts in particular are more affected than others
- it is noticeable that many of these districts were characterized by a lively life before Corona
- this research looks at whether there are certain neighborhoods that are at particularly high risk of Corona due to their venue structure



Data

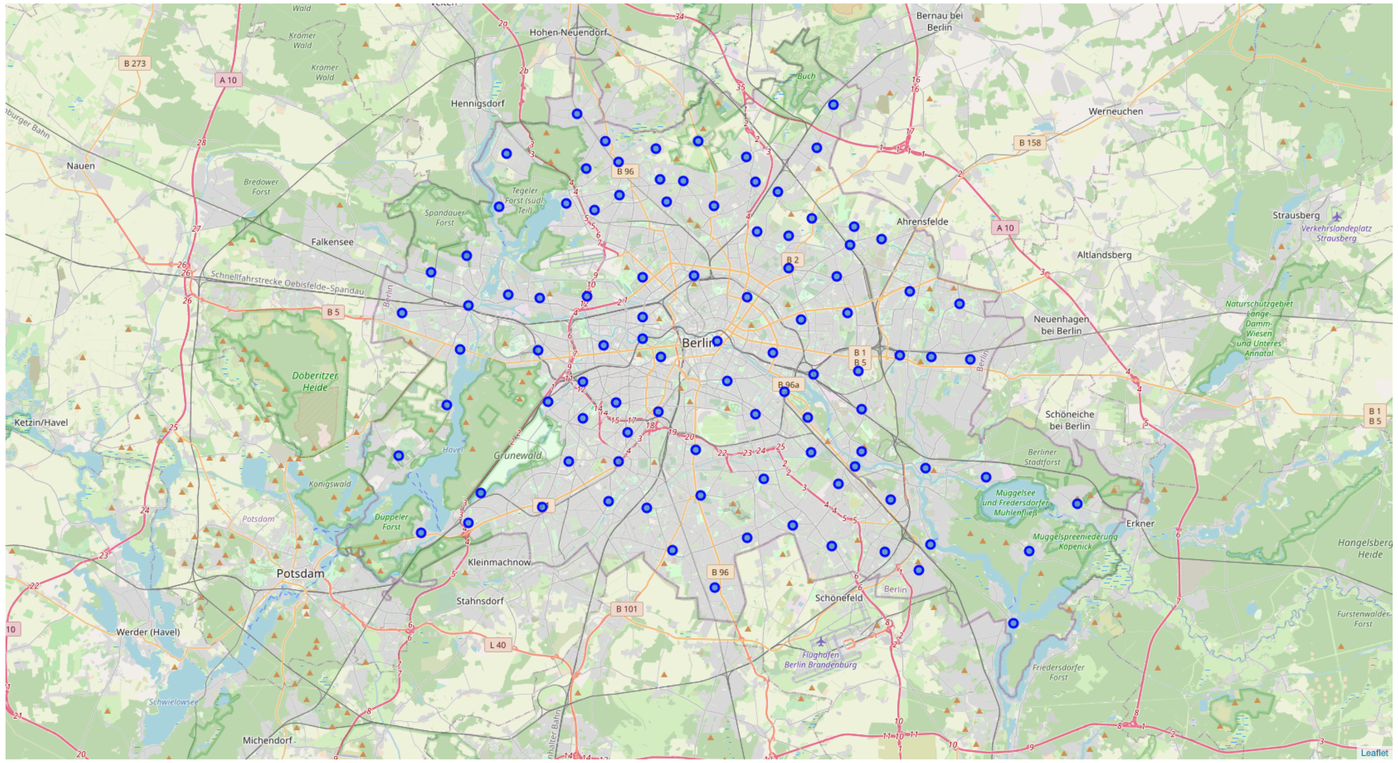
The following data sources were used to investigate the problem

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

Method

Berlin 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.



Districts of Berlin

Method

Berlin Venues

- These coordinates of the districts were 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

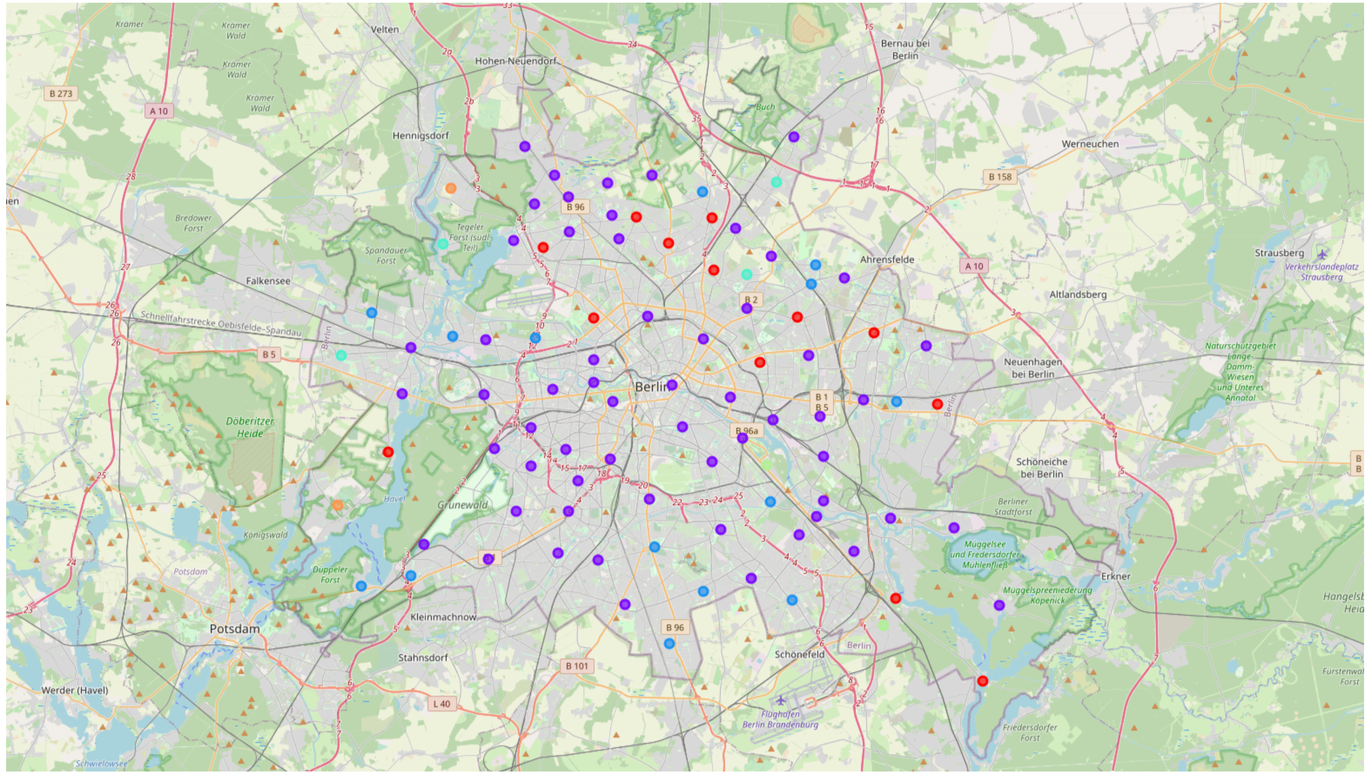
Mitte	Alte Nationalgalerie	52.520796	13.398673	Art Museum
Schöneberg	Alte Schmiede bei Pino	52.482138	13.357734	Italian Restaurant
Dahlem	Alter Krug Dahlem	52.457550	13.288223	German Restaurant
Tempelhof	Alter Park	52.462871	13.382309	Park
Mitte	Altes Museum	52.519537	13.398803	History Museum
Gesundbrunnen	Altin Saray	52.551748	13.382577	Turkish Restaurant
Tempelhof	Amera	52.467170	13.382752	Italian Restaurant
Märkisches Viertel	American Western Saloon	52.597617	13.352591	American Restaurant
Moabit	Amstel House	52.528478	13.336745	Hostel

Excerpt from the list of districts, venues and geo data

Method

Clustering districts according to venues

- 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.



Berlin districts colored according to their assigned cluster

Method

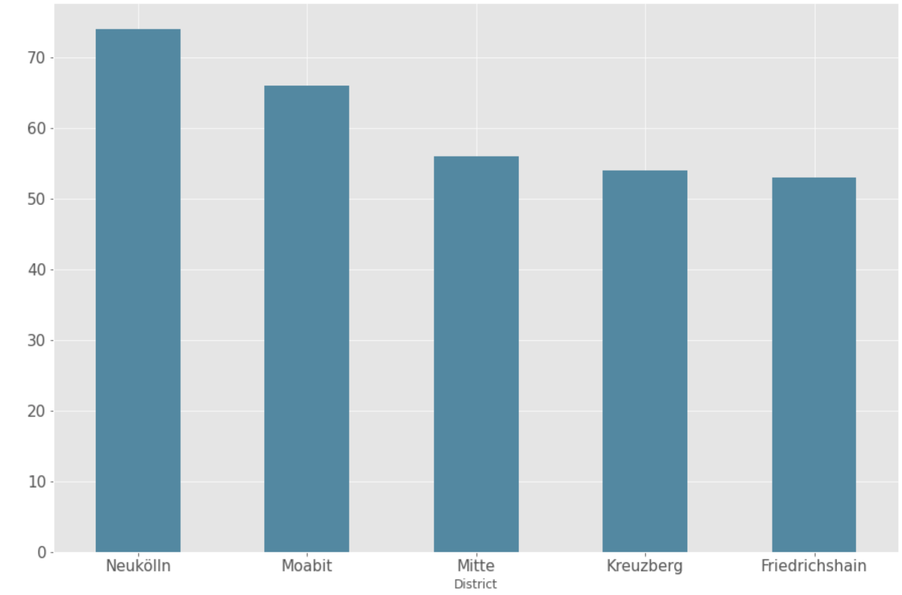
Description of the Clusters

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

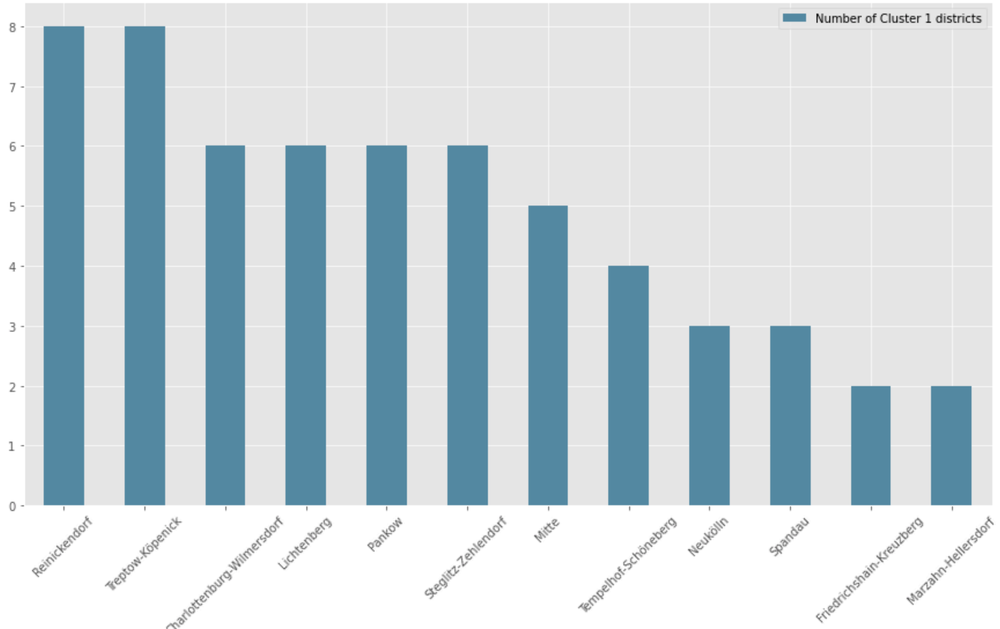
Method

Infection numbers

- 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.



Berlin districts with the highest number of venues



Berlin districst with the highest Corona incidence

Hypotheses

2 hypotheses were tested

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

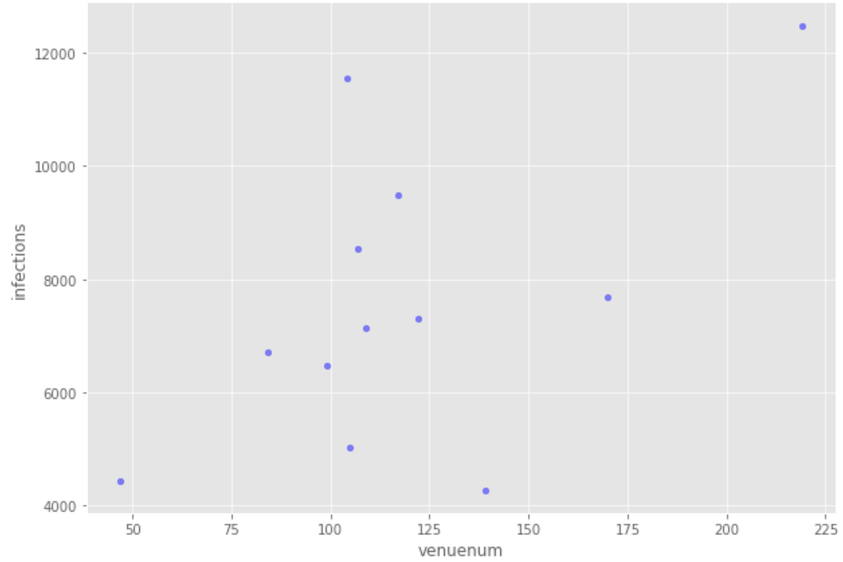
To examine possible relations with Corona infections, a regression analysis was performed.

Results

No statistical evidence for the hypotheses

- There is a positive correlation between the total number of venues in a borough and the incidence of covid infections. (p = .38, alpha = .05)

- There is a positive correlation between the number of "vibrant" clusters in a borough and the incidence of covid infections. (p=.06, alpha = .05)



Scatterplot with the data points for numbers of venues and infections in all boroughs.

A tendency towards a positive relationship is visible, but the result of the regression analysis showed no statistically significant correlation.

Discussion

Limitations to this investigation

- 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.

Discussion

Limitations to this investigation

- 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.