Coronavirus pandemic in Berlin: When should the lockdown end?

Efrem Gebremicael

April 28, 2020

1. Introduction

1.1 Background

Until today almost three million coronavirus cases confirmed in 185 countries, with more than 205,000 deaths, according to data compiled by Johns Hopkins University. The virus was first detected in the city of Wuhan, China, in late 2019. COVID-19 is the name given to the disease associated with the virus. The US has by far the largest number of cases, with more than 900,000 confirmed infections. In Europe, the counties with the most confirmed infections are Spain, Italy, France, Germany and the UK. But fortunately, we in Germany, have less than the half of the recorded death in Spain, Italy, France and the UK (all recorded more than 20,000 deaths).

1.2 Problem

The coronavirus is spreading rapidly in many countries and the number of deaths is still climbing, also in Berlin. Berlin is the capital and largest city of Germany by both area (892 square kilometers) and population (around 3.6 million). Lockdowns have been essential for containing the spread of coronavirus, but they have huge negative effect on the economy and the population well-being. Where, when and which restrictions to lift, are big decisions for governments to make.

This project could contribute to the decision-making process. I will analyze if there is a relationship between population density in the different boroughs in Berlin and confirmed infections. I will also use location data to show if boroughs with the hottest places (i.e. many people in an area) have the highest confirmed coronavirus cases.

2. Data Description

I used the official website of Berlin to get the actual confirmed coronavirus cases in the different boroughs and for the Berlin boroughs data (including Density per km²), a Wikipedia page exists that has all the information I needed. I just needed to use the pandas.read_html function to read the data into a pandas dataframe. After some wrangling and cleaning, the data was in a structured format.

For creating Choropleth maps, I needed a geojson file. After a Google search, I found a Berlin boroughs geojson file on GitHub. For the location data, I used the Foursquare API to get the recommended venues and top picks in Berlin. The location data was used to create markers in the maps.

3. Methodology - Exploratory data analysis

3.1 Creating a dataset

In this study the focus was on explanatory data analysis. I, as probably many other people, thought that there is a relationship between population density and confirmed coronavirus cases. Explanatory data analysis is helpful to check if this assumption is correct.

In first step I have downloaded the required data and after some wrangling and cleaning, I had a dataset for further analysis (Figure 1).

	Borough	Density per km²	COVID-19 Count
0	Mitte	9733.0	861.0
1	Friedrichshain-Kreuzberg	14246.0	460.0
2	Pankow	3956.0	600.0
3	Charlottenburg-Wilmersdorf	5289.0	672.0
4	Spandau	2656.0	236.0
5	Steglitz-Zehlendorf	3010.0	485.0
6	Tempelhof-Schöneberg	6622.0	587.0
7	Neukölln	7338.0	632.0
8	Treptow-Köpenick	1610.0	307.0
9	Marzahn-Hellersdorf	4347.0	289.0
10	Lichtenberg	5592.0	241.0
11	Reinickendorf	2970.0	451.0

Figure 1. Dataset with *Borough, Density per km², COVID-19 Count*

3.2 Scatter plot and correlation calculation

I have started with a scatter plot (Figure 2) to see if we can spot a relationship between population density in the different boroughs in Berlin and confirmed coronavirus cases. I observed no clear relationship between Density per km² and COVID-19 Count.

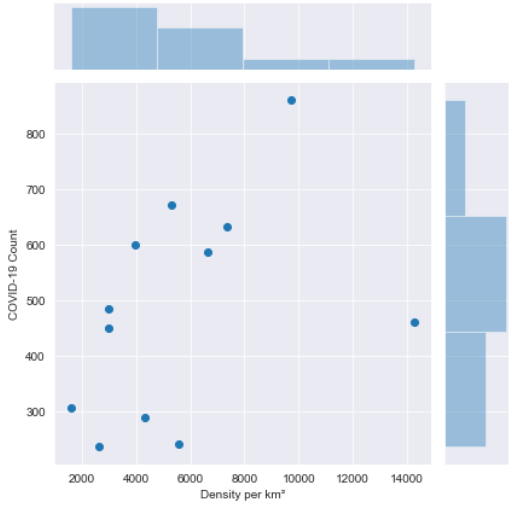

Figure 2. Relationship between *Density per km²* and *COVID-19 Count*

I also made a Pandas correlation (Pearson) calculation and it proves that we have no significant correlation (0.426).

3.3 Clustering

I have used *k*-means Clustering to segment the Berlin boroughs. I took the advantage of the Elbow method (Figure 3) and decided to make 5 clusters (Figure 4). The cluster plot shows that the green and red clusters contain a single borough. The green cluster is the borough Friedrichshain-Kreuzberg with the highest population density, but moderate infections and the red cluster is Mitte with the highest coronavirus cases (Figure 5). The orange cluster, including Treptow-Köpenick, Marzahn-Hellersdorf, Lichtenberg and Spandau, has low infections. The blue cluster with Pankow, Steglitz-Zehlendorf and Reinickendorf has moderate coronavirus cases and the purple cluster with Charlottenburg-Wilmersdorf, Tempelhof-Schöneberg and Neukölln has a bit more infections than the blue cluster.

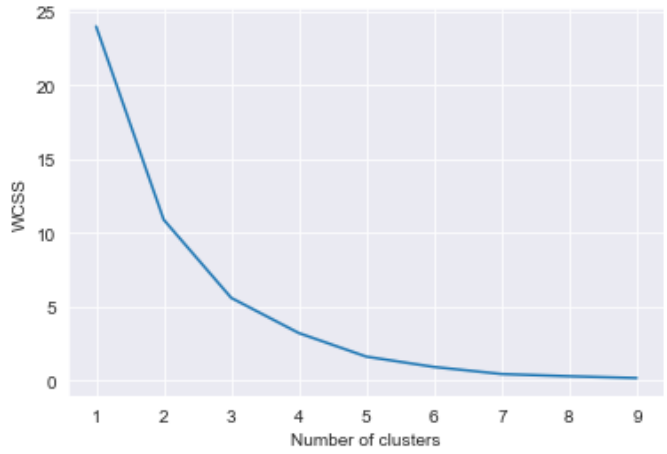

Figure 3. Number of clusters and Within-Cluster-Sum-of-Squares

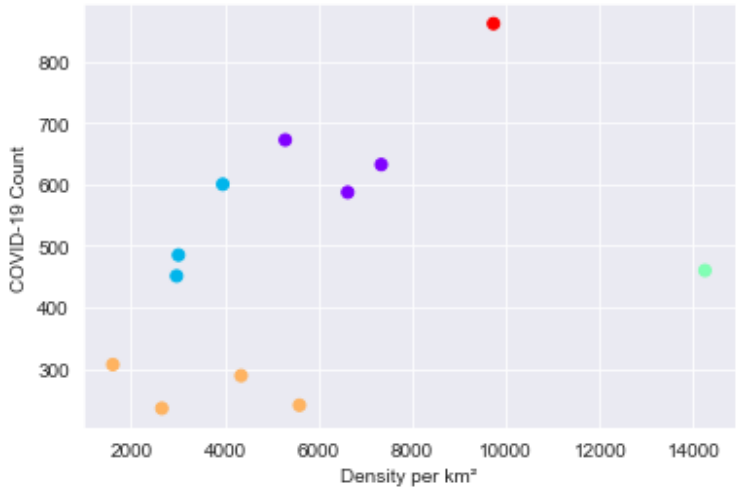

Figure 4. 5-clusters. The green cluster is the borough Friedrichshain-Kreuzberg, the red cluster is Mitte, the orange cluster includes Treptow-Köpenick, Marzahn-Hellersdorf, Lichtenberg and Spandau, the blue cluster Pankow, Steglitz-Zehlendorf and Reinickendorf and the purple cluster Charlottenburg-Wilmersdorf, Tempelhof-Schöneberg and Neukölln

	Borough	Density per km²	COVID-19 Count	cluster_pred
0	Mitte	9733.0	861.0	4
1	Friedrichshain-Kreuzberg	14246.0	460.0	2
2	Pankow	3956.0	600.0	0
3	Charlottenburg-Wilmersdorf	5289.0	672.0	1
4	Spandau	2656.0	236.0	3
5	Steglitz-Zehlendorf	3010.0	485.0	0
6	Tempelhof-Schöneberg	6622.0	587.0	1
7	Neukölln	7338.0	632.0	1
8	Treptow-Köpenick	1610.0	307.0	3
9	Marzahn-Hellersdorf	4347.0	289.0	3
10	Lichtenberg	5592.0	241.0	3
11	Reinickendorf	2970.0	451.0	0

Figure 5. Data frame with the predicted clusters

3.4 Visualize corona cases and venues

I also used location data and the Python library Folium to visualize if boroughs with the hottest places have the highest confirmed infections. My assumption was that boroughs with many recommended venues are crowded with people and therefore the spread of the coronavirus is more likely.

I started with a Choropleth map without markers to visualize infections in the boroughs and then I include markers using the Foursquare API to get the location data of recommended places. I first used the property *topPicks* (a mix of recommendations generated without a query from the user) (Figure 6) and then I searched for recommended venues without specifying any category or property (Figure 7). I observed, in both maps with markers, no relation between boroughs with many recommended venues and boroughs with high confirmed coronavirus cases.

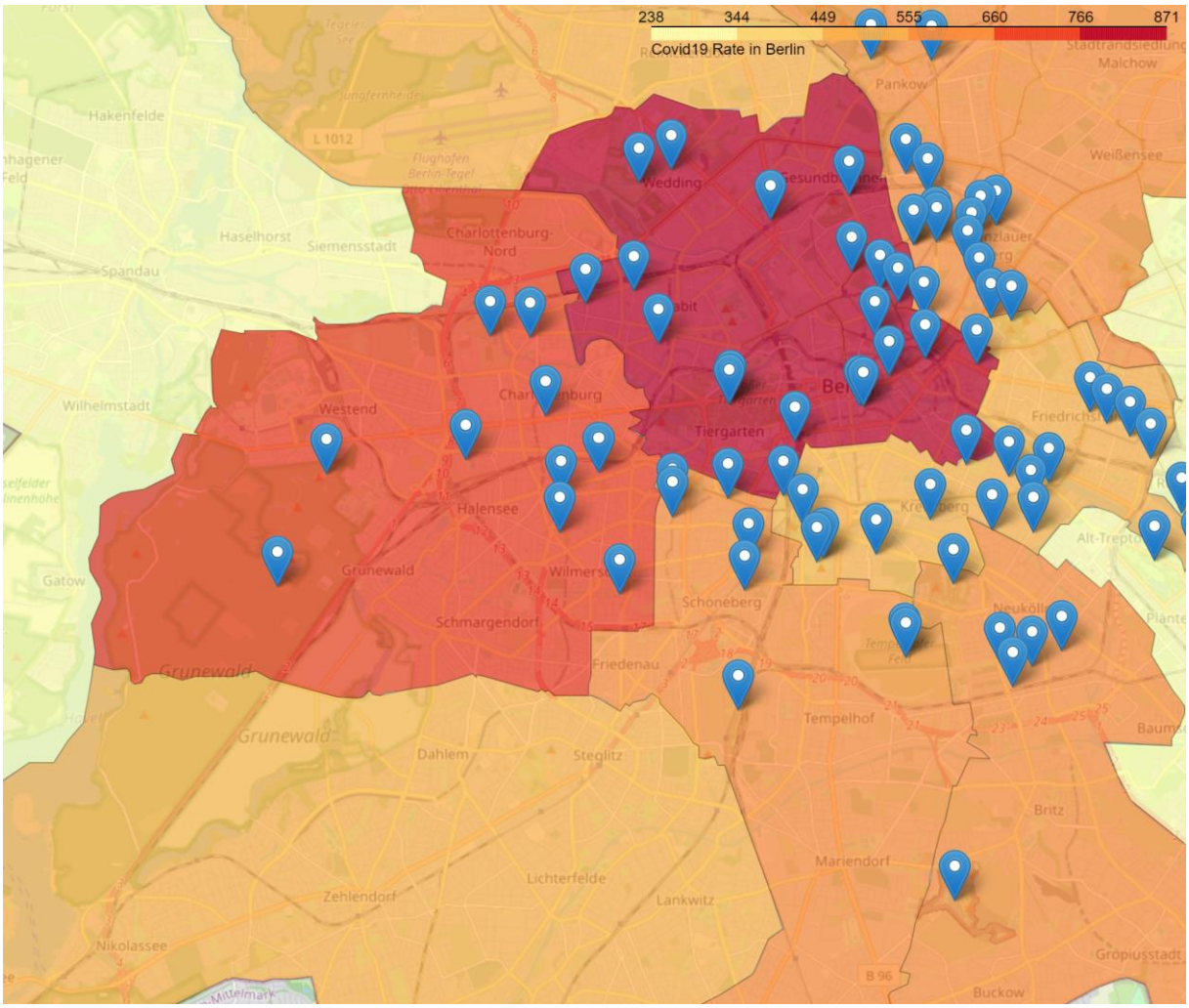

Figure 6. Choropleth map with top picks markers

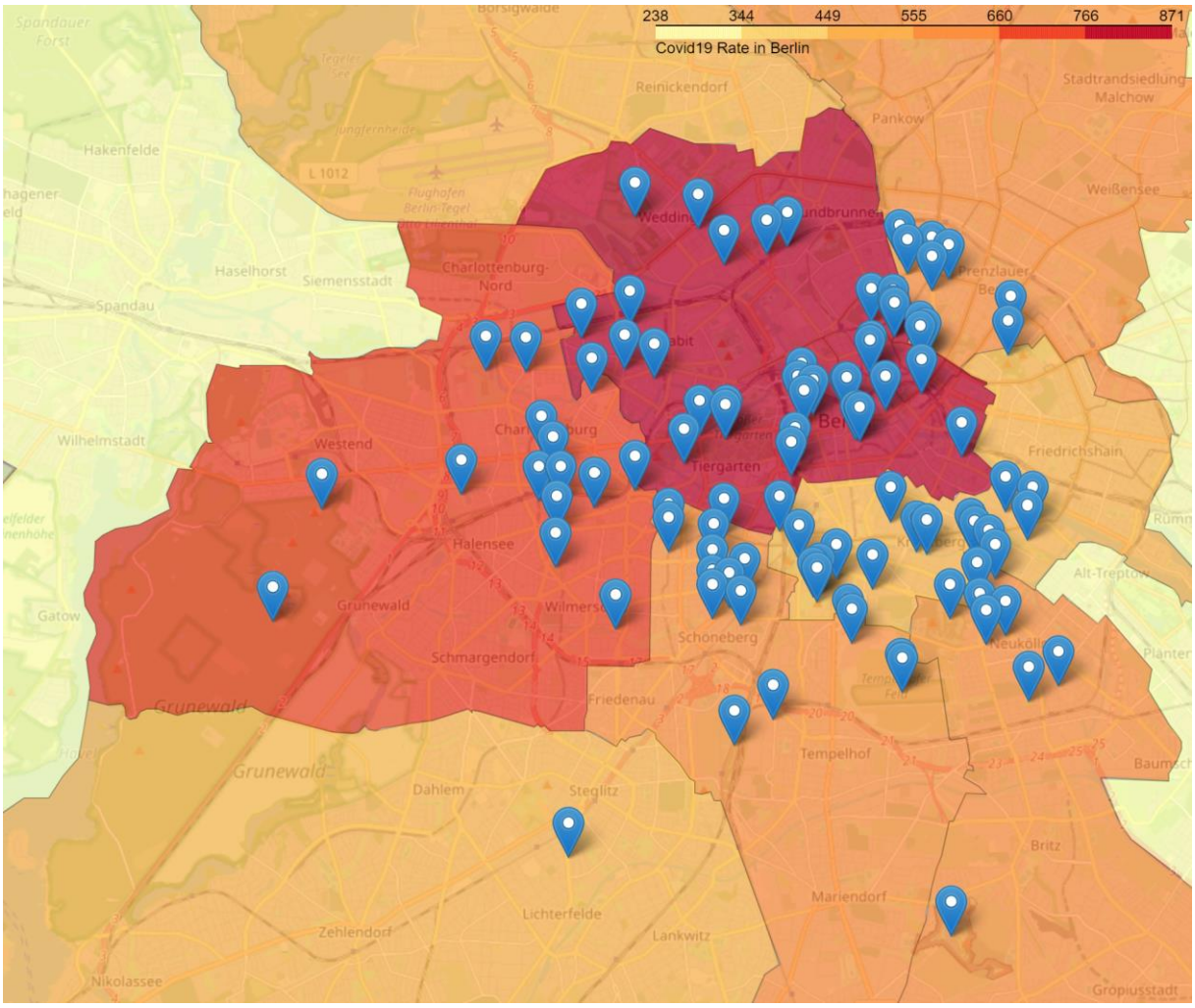

Figure 7. Choropleth map with recommended venues markers

4. Results and Discussion

This project started with the question “when should the lockdown end”. Where, when and which restrictions to lift, are big decisions for governments to make. Our government must balance saving lives today with long-term damage to society. The risk of lifting the lockdown to soon, is another explosive outbreak.

This study shows that these decisions are indeed very difficult. In an Explanatory data analysis, I observed no clear relationship between population density in the different Berlin boroughs and confirmed coronavirus cases. For example, the borough Friedrichshain-Kreuzberg has the highest population density but moderate infections (461 until today). The borough Mitte has the highest coronavirus cases (861 until today) but the second highest population density. Simply start lifting restrictions according to population density will not work, because as a result, restrictions are lifted in Mitte with the highest coronavirus cases before the restrictions are lifted in Friedrichshain-Kreuzberg.

The Choropleth maps with makers in this project shows that broughs with many recommended venues does not implicate high spread of the coronavirus. For example, the borough Friedrichshain-Kreuzberg has moderate coronavirus cases although the borough is trendy, popular among tourists and Berliners and offers many recommended venues. Therefore, we can hope that we have not another high spread of coronavirus when the venues are open again.

5. Conclusion

In this study I used Explanatory data analysis to discover insight from Berlin borough data and location data. I observed no clear relationship between population density in the different Berlin boroughs and confirmed coronavirus cases. Also, broughs with many recommended venues does not implicate high coronavirus infections. Further and continuous analysis with more data (e.g. including demographics, more tested people) is necessary to discover better insight and as a result better decision. I think the strategy of our government, bringing cases down quickly and making every two weeks decisions regarding the restrictions, is a good plan.