Applied Data Science Capstone Project – How can we better prevent Covid-19?

Prologue

In the recent months, the second wave of Covid-19 is coming worldwide, millions of people are getting involved in this crise.
After graduation, I have plenty of time finding a job, thanks to covid-19 :(So to use my time meaningfully, I decided to learn the IBM data science course on Coursera. This project is the final assignment of this Coursera course, instead of finishing the boring task from this course, I decided to do something “in real”, perhaps with the power of data science and machine learning, I can grab some idea helping me and the public getting rid of Covid-19.
I have been in Germany for 5 years, Germany is like a second home for me. Due to Covid-19 a new lock-down just happened here. But what should we do after this lock-down? How can we celebrate X’mas with friends in a public area? How can we prevent the third wave of covid-19?

Introduction/Business problem

Berlin, the capital city of Germany, attracts a millions of tourists each year, as a traveler in Berlin under the covid-19, you may ask which place should I go / don’t go, is the chance of catching covid-19 in a germen restaurant lower than a Chinese restaurant? Will, you may find out the answer in this project.
As requested by the final assignment task from IBM data science, I’ll use the foursquare data from Folium API to show the covid-19 related data in Berlin neighborhood: neighborhoods name, ID, location and venues category etc.
One more thing, I got the overview of districts of Berlin via Wikipedia: https://de.wikipedia.org/wiki/Berlin. With this link you will get a table with all city districts.

Eckdaten der Bezirke von Berlin am 31. Dezember 2019[42][4]

Nr.	Bezirk von Berlin	Einwohner [Anm. 1]	Fläche in km²	Einwohner pro km²
1.	Mitte	385.748	39,47	9.733
2.	Friedrichshain-Kreuzberg	290.386	20,34	14.246
3.	Pankow	409.335	103,07	3.956
4.	Charlottenburg-Wilmersdorf	343.592	64,72	5.289
5.	Spandau	245.197	91,87	2.656
6.	Steglitz-Zehlendorf	310.071	102,56	3.010
7.	Tempelhof-Schöneberg	350.984	53,10	6.622
8.	Neukölln	329.917	44,93	7.338
9.	Treptow-Köpenick	273.689	168,42	1.610
10.	Marzahn-Hellersdorf	269.967	61,78	4.347
11.	Lichtenberg	294.201	52,12	5.592
12.	Reinickendorf	266.408	89,31	2.970
	Land Berlin (gesamt)	3.669.491	891,68	4.088

I will use those districts and the data about public facilities such as bar, restaurant, museum and supermarket etc.

Methodology

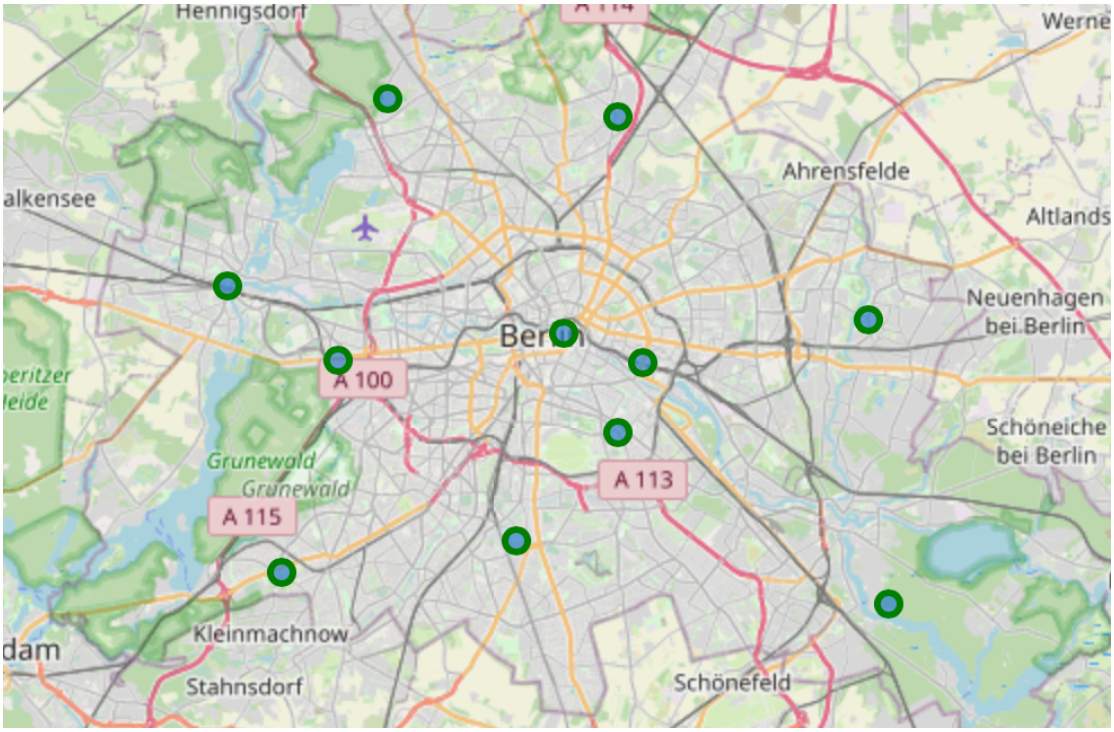
Because of the germen data protection regulation ”Datenschutz”, it is really hard to get detailed dataset about covid-19 in the scope of city district in Berlin. So, I reconsidered to plot the most visited venues in berlin that sorted by city districts, and cluster them into 5 clusters, to suggest the public which place they may avoid, but Germany is a free country, so the suggestions of don’t-go-area also contain a don’t-go-business-location such as supermarket, coffee shop, restaurant etc.

In this section I’ll use the data about venues in Berlin city to plot the dangerous area with 5 different clusters, and visualize the data with cluster map and heat map.

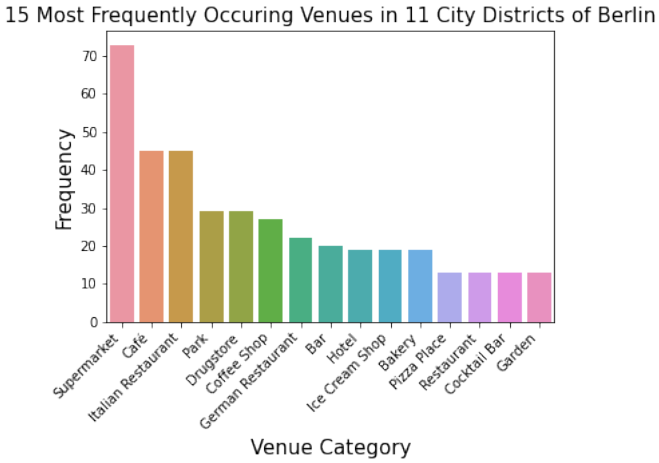
In the beginning, I scraped the data from Wiki and combine the city locations with it’s represents latitude and longitude and store the values in a data frame. Results see table below.

展开输出; 双击隐藏	Borough	Population	Area	Density	Latitude	Longitude
0	Charlottenburg-Wilmersdorf	319628	64.72	4878	52.507856	13.263952
1	Friedrichshain-Kreuzberg	268225	20.16	13187	52.506862	13.450642
2	Lichtenberg	259881	52.29	4952	48.921296	7.481227
3	Marzahn-Hellersdorf	248264	61.74	4046	52.522523	13.587663
4	Mitte	332919	39.47	8272	52.517690	13.402376
5	Neukölln	310283	44.93	6804	52.481150	13.435350
6	Pankow	366441	103.01	3476	52.597917	13.435316
7	Reinickendorf	240454	89.46	2712	52.604763	13.295287
8	Spandau	223962	91.91	2441	52.535788	13.197792
9	Steglitz-Zehlendorf	293989	102.50	2818	52.429205	13.229974
10	Tempelhof-Schöneberg	335060	53.09	6256	52.440603	13.373703
11	Treptow-Köpenick	241335	168.42	1406	52.417893	13.600185

After data preprocessing, I can easily plot the 12 city districts (boroughs) on map by using the Folium API:



After that I just retrieved the foursquare data for all venues on Foursquare with radius of 3000 meters. Each green dot from the chart above represents a centroid, and the Folium API with search all the business location within 3000 meters based on those 12 centroid. With the above search criteria as input, the Folium returned a data frame with shape(947,7) which means 947 business location were found.
Then, I plotted the result with a bar chart by using the seaborn and the matplotlib.pyplot API:



The bar chart above showed the frequency of the top-15 business location in Berlin city, which suggests the top15 don't-go-business-location with covid-19.

Now it's time for Clustering. To find clusters of business locations in different city districts by one-hot encoding, I transformed the data frame with the restaurant venues with neighborhoods as index. The one-hot encoding returned a data frame with shape (947, 214):

	Neighborhoods	ATM	Adult Boutique	African Restaurant	American Restaurant	Amphitheater	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	BBQ Joint	Bagel Shop	Bakery
0	Charlottenburg-Wilmersdorf	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Charlottenburg-Wilmersdorf	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Charlottenburg-Wilmersdorf	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Charlottenburg-Wilmersdorf	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Charlottenburg-Wilmersdorf	0	0	0	0	0	0	0	0	0	0	0	0	0	0

```
1 berlin_onehot1.shape
```

```
(947, 214)
```

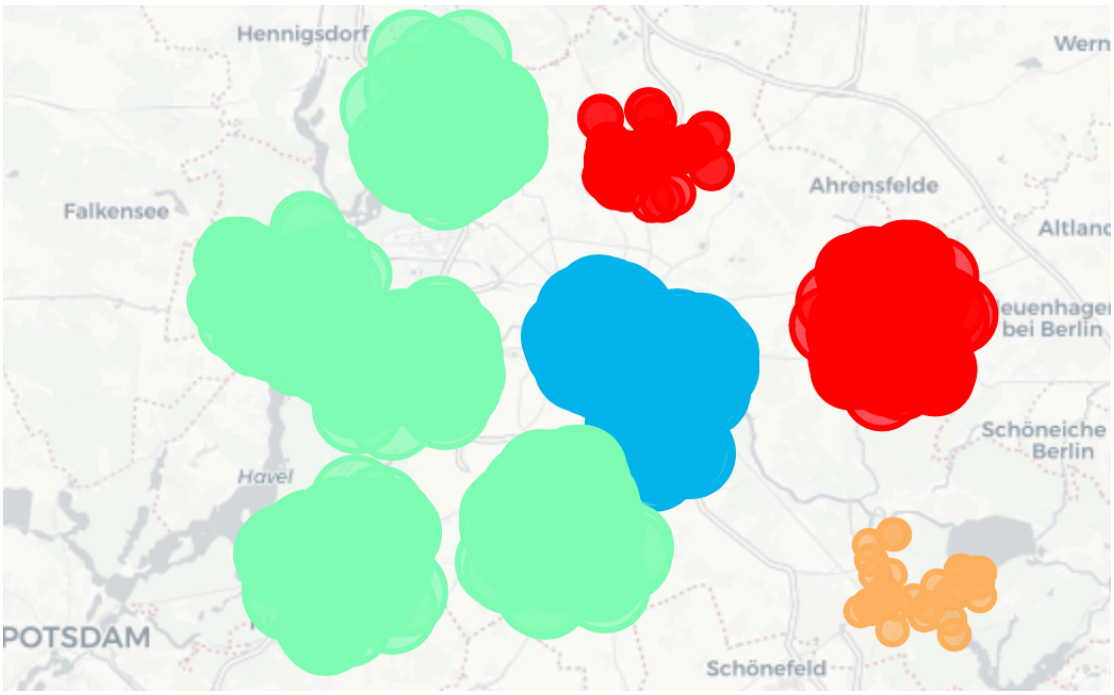
Next, I grouped the data frame by city districts to show the frequency of each business locations:

	Neighborhoods	ATM	Adult Boutique	African Restaurant	American Restaurant	Amphitheater	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	BBQ Joint	Bagel Shop	B
0	Charlottenburg-Wilmersdorf	0.000000	0.00	0.00	0.000000	0.01	0.010000	0.00	0.03	0.00	0.020000	0.010000	0.00	0.00	0.0
1	Friedrichshain-Kreuzberg	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.00	0.00	0.00	0.000000	0.000000	0.01	0.01	0.0
2	Lichtenberg	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.00	0.00	0.00	0.000000	0.000000	0.00	0.00	0.0
3	Marzahn-Hellersdorf	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.00	0.00	0.00	0.011628	0.000000	0.00	0.00	0.0
4	Mitte	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.01	0.01	0.01	0.000000	0.000000	0.01	0.00	0.0
5	Neukölln	0.000000	0.00	0.03	0.000000	0.00	0.000000	0.00	0.00	0.00	0.000000	0.000000	0.00	0.00	0.0
6	Pankow	0.021277	0.00	0.00	0.000000	0.00	0.000000	0.00	0.00	0.00	0.021277	0.000000	0.00	0.00	0.0
7	Reinickendorf	0.000000	0.01	0.00	0.000000	0.00	0.010000	0.00	0.00	0.00	0.000000	0.010000	0.00	0.00	0.0
8	Spandau	0.000000	0.00	0.00	0.000000	0.00	0.025316	0.00	0.00	0.00	0.012658	0.012658	0.00	0.00	0.0
9	Steglitz-Zehlendorf	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.00	0.00	0.00	0.020000	0.000000	0.00	0.02	0.0
10	Tempelhof-Schöneberg	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.00	0.00	0.00	0.010000	0.010000	0.00	0.00	0.0
11	Treptow-Köpenick	0.000000	0.00	0.00	0.030303	0.00	0.000000	0.00	0.00	0.00	0.000000	0.000000	0.00	0.00	0.0

Well, the data frame above is a little bit hard for human to read. So I create another data frame for human ☺

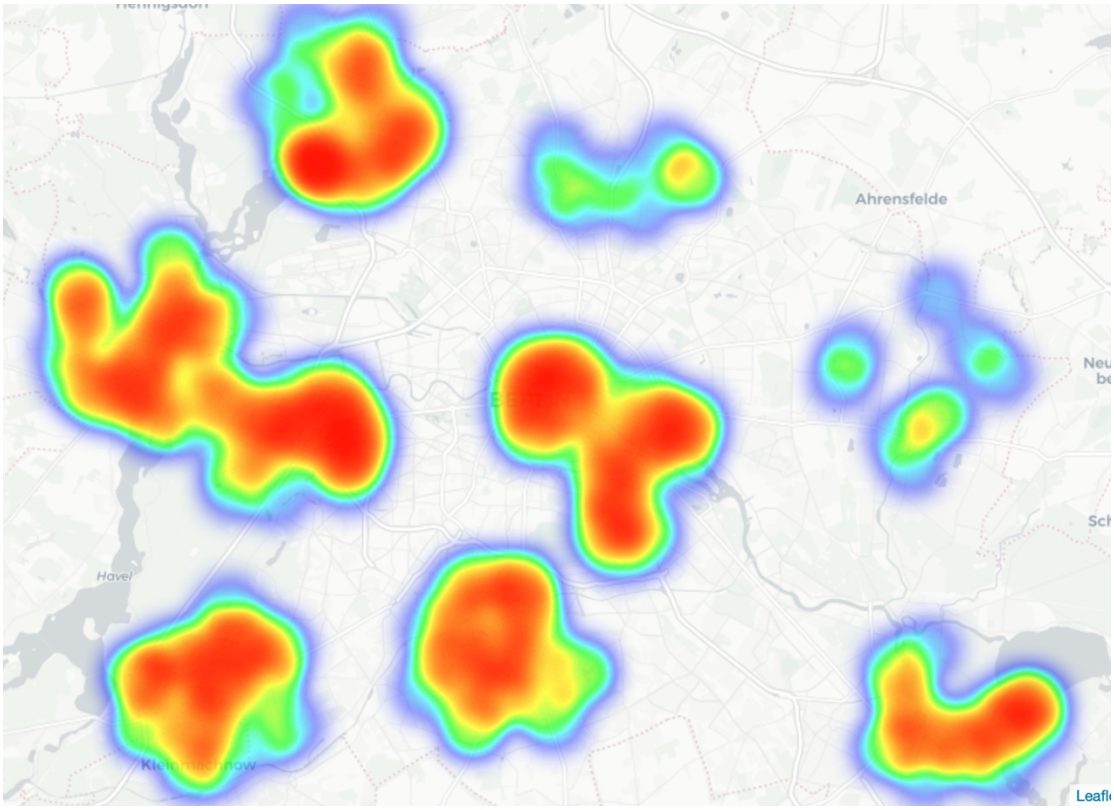
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	11th Most Common Venue	12th M Comm Ver
0	Charlottenburg-Wilmersdorf	Italian Restaurant	Café	German Restaurant	Hotel	Trattoria/Osteria	Gym / Fitness Center	Vietnamese Restaurant	Art Museum	Soccer Stadium	Supermarket	Ice Cream Shop	
1	Friedrichshain-Kreuzberg	Falafel Restaurant	Coffee Shop	Bar	Café	Ice Cream Shop	Cocktail Bar	Brewery	Italian Restaurant	Pizza Place	Nightclub	Thai Restaurant	Vegetar / Veg Restaur
2	Lichtenberg	Hostel	Historic Site	Yoga Studio	Drugstore	Fish & Chips Shop	Field	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Exhibit	Event Space	Electron St
3	Marzahn-Hellersdorf	Supermarket	Garden	Drugstore	Shopping Mall	Big Box Store	Park	Fast Food Restaurant	Tram Station	Bakery	Ice Cream Shop	Café	Gy Fitn Cer
4	Mitte	Coffee Shop	Bookstore	Café	Hotel	Park	Beer Bar	Sandwich Place	Italian Restaurant	Monument / Landmark	Middle Eastern Restaurant	Concert Hall	Thea

Finally, I could cluster this data frame by using k-means and plot them on a map:



This map contains basically only 4 clusters, I don’t know why but the last cluster is in France. ☹

Using the same parameter, I plotted a heatmap as well:



From the heatmap above we can discover the dangerous city parts of berlin which belong to the don’t go place.

Next, I’m going to describe each area by the cluster result:

Result

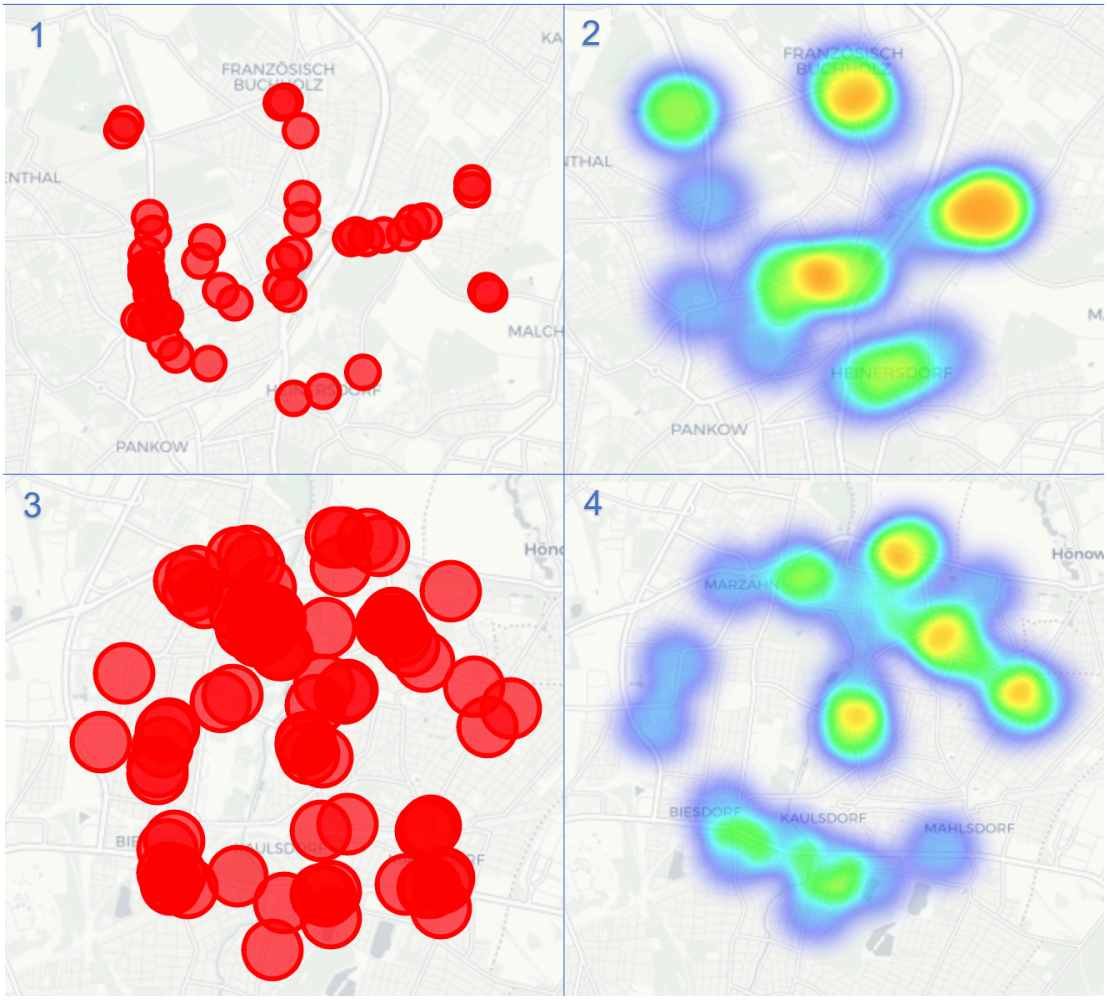
This is the overview of each cluster sorted with each city district:

	Berlin Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	3	Charlottenburg-Wilmersdorf	Italian Restaurant	Café	German Restaurant	Hotel	Trattoria/Osteria	Gym / Fitness Center	Vietnamese Restaurant	Art Museum	Soccer Stadium	Supermarket
1	2	Friedrichshain-Kreuzberg	Falafel Restaurant	Coffee Shop	Bar	Café	Ice Cream Shop	Cocktail Bar	Brewery	Italian Restaurant	Pizza Place	Nightclub
2	1	Lichtenberg	Hostel	Historic Site	Yoga Studio	Drugstore	Fish & Chips Shop	Field	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Exhibit
3	0	Marzahn-Hellersdorf	Supermarket	Garden	Drugstore	Shopping Mall	Big Box Store	Park	Fast Food Restaurant	Tram Station	Bakery	Ice Cream Shop
4	2	Mitte	Coffee Shop	Bookstore	Café	Hotel	Park	Beer Bar	Sandwich Place	Italian Restaurant	Monument / Landmark	Middle Eastern Restaurant
5	2	Neukölln	Coffee Shop	Bar	Café	Cocktail Bar	Vegetarian / Vegan Restaurant	Italian Restaurant	Indie Movie Theater	Ice Cream Shop	African Restaurant	Pizza Place
6	0	Pankow	Supermarket	Italian Restaurant	Bus Stop	Park	Greek Restaurant	Hotel	Drugstore	Café	Tram Station	Bakery
7	3	Reinickendorf	Supermarket	Italian Restaurant	Drugstore	German Restaurant	Restaurant	Clothing Store	Indian Restaurant	Fast Food Restaurant	Motorcycle Shop	Café
8	3	Spandau	Supermarket	Bus Stop	Drugstore	German Restaurant	Pizza Place	Park	Ice Cream Shop	Trattoria/Osteria	Restaurant	Big Box Store
9	3	Steglitz-Zehlendorf	Café	Italian Restaurant	Supermarket	German Restaurant	Drugstore	Gas Station	Organic Grocery	Park	Doner Restaurant	Hotel
10	3	Tempelhof-Schöneberg	Supermarket	Park	Italian Restaurant	Drugstore	Café	Pool	Bakery	Ice Cream Shop	Taverna	Doner Restaurant
11	4	Treptow-Köpenick	Supermarket	Beach	Hotel	Mountain	Historic Site	Shopping Mall	Forest	Boat Rental	Bowling Alley	Scenic Lookout

Cluster 0:

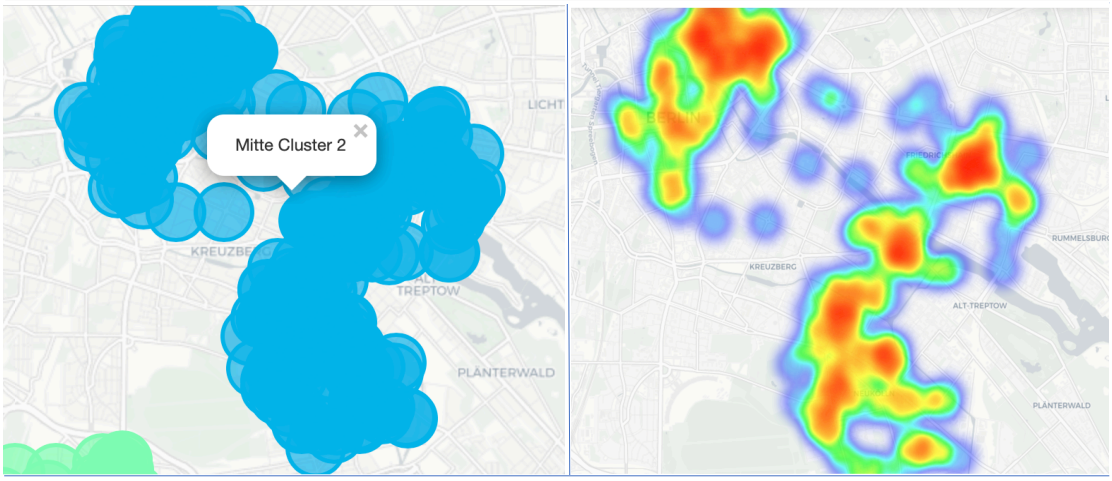
Cluster 0 contains 2 neighborhoods: Marzahn-Hellersdorf and Pankow,

	Neighborhood	Berlin Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
3	Marzahn-Hellersdorf	0	Marzahn-Hellersdorf	Supermarket	Garden	Drugstore	Shopping Mall	Big Box Store	Park	Fast Food Restaurant	Tram Station	Bakery	Ice Cream Shop
6	Pankow	0	Pankow	Supermarket	Italian Restaurant	Bus Stop	Park	Greek Restaurant	Hotel	Drugstore	Café	Tram Station	Bakery



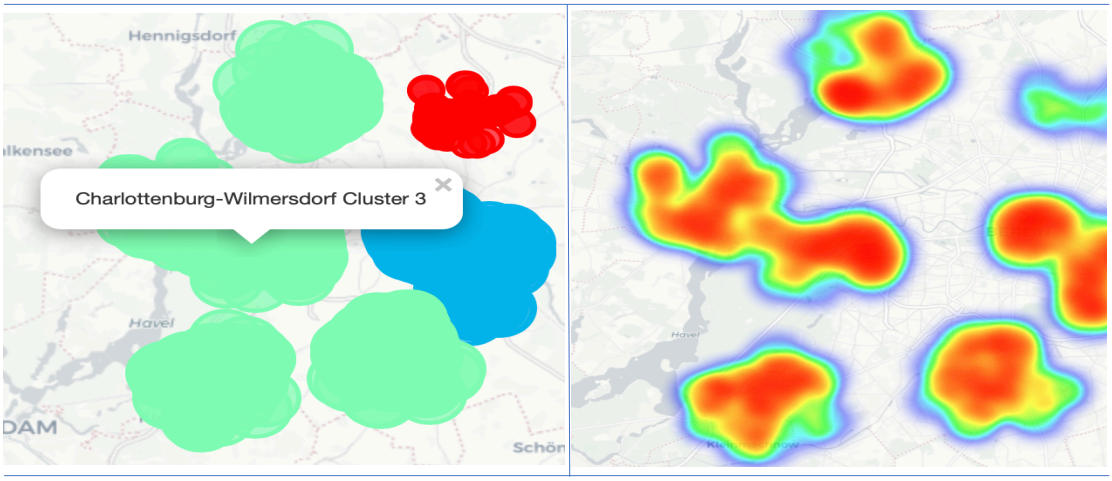
Cluster 2:

	Neighborhood	Berlin Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Friedrichshain-Kreuzberg	2	Friedrichshain-Kreuzberg	Falafel Restaurant	Coffee Shop	Bar	Café	Ice Cream Shop	Cocktail Bar	Brewery	Italian Restaurant	Pizza Place	Nightclub
4	Mitte	2	Mitte	Coffee Shop	Bookstore	Café	Hotel	Park	Beer Bar	Sandwich Place	Italian Restaurant	Monument / Landmark	Middle Eastern Restaurant
5	Neukölln	2	Neukölln	Coffee Shop	Bar	Café	Cocktail Bar	Vegetarian / Vegan Restaurant	Italian Restaurant	Indie Movie Theater	Ice Cream Shop	African Restaurant	Pizza Place



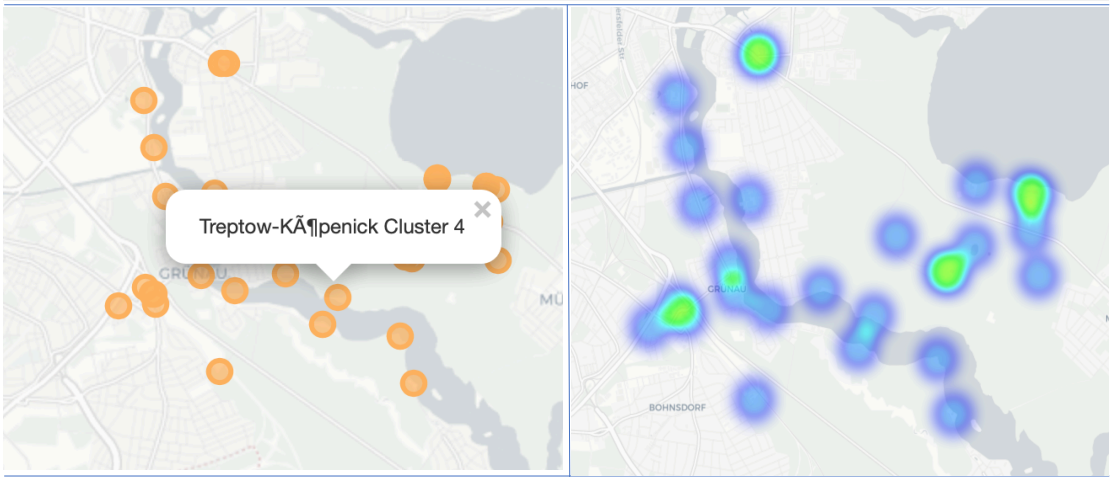
Cluster 3:

	Neighborhood	Berlin Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
0	Charlottenburg-Wilmersdorf	3	Charlottenburg-Wilmersdorf	Italian Restaurant	Café	German Restaurant	Hotel	Trattoria/Osteria	Gym / Fitness Center	Vietnamese Restaurant	Art Museum	Soccer Stadium
7	Reinickendorf	3	Reinickendorf	Supermarket	Italian Restaurant	Drugstore	German Restaurant	Restaurant	Clothing Store	Indian Restaurant	Fast Food Restaurant	Motorcycle Shop
8	Spandau	3	Spandau	Supermarket	Bus Stop	Drugstore	German Restaurant	Pizza Place	Park	Ice Cream Shop	Trattoria/Osteria	Restaurant
9	Steglitz-Zehlendorf	3	Steglitz-Zehlendorf	Café	Italian Restaurant	Supermarket	German Restaurant	Drugstore	Gas Station	Organic Grocery	Park	Doner Restaurant
10	Tempelhof-Schöneberg	3	Tempelhof-Schöneberg	Supermarket	Park	Italian Restaurant	Drugstore	Café	Pool	Bakery	Ice Cream Shop	Taverna



Cluster 4:

	Neighborhood	Berlin Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
11	Treptow-Köpenick	4	Treptow-Köpenick	Supermarket	Beach	Hotel	Mountain	Historic Site	Shopping Mall	Forest	Boat Rental	Bowling Alley	Scenic Lookout



The Cluster 1 is out of the scope of Berlin.

After displaying the cluster maps and heatmaps, we could get some conclusions, for example in Berlin Mitte, city center, we should avoid street with coffee shop to prevent covid-19 and from the heat map of cluster 1 (Berlin Mitte), the area with red mark should also be avoided.

Cluster 4 (Berlin Treptow-Koepenick) represents a safer place, for those who can't bear the lock down, can perhaps go there and meet some friends but be careful!

At last, I have to answer the question: is the chance of catching covid-19 in a germen restaurant lower than a Chinese restaurant? Well, the answer is negative, because the most germen restaurants are in the busy area in Berlin, so the chance of catching covid-19 is even higher.

Discussion

From this project I obtained the abilities in different Python APIs such as Matplotlib, Pandas, Seaborn, Folium etc. For data visualization technique, I have learned how to preprocess datasets from internet source and transform them into a desired shape to build the model. The results of K-means Clustering is astonishing, the machine learning algorithm behind this scene can find so many hidden patterns. Sometimes I feel pathetic being a human, because the datasets are right there, but I can't find any interrelations from them, and it hurts.

Thanks to Github, different ideas are open-sourced, and they just inspired me to finish this project. I just realized that there is still a long way to go after finishing this course.

Until then, stay tuned and enjoy AI!

Conclusion

I have used the power of machine leaning and data science to achieve the goal of

preventing getting a covid-19 in Berlin as a tourist, so if you are planning to travel there or already there, please feel free the check the heatmap, if you could get some useful information from it, I'll be happy.
One more thing about this project: The methods I used is only based on the number of restaurant, there is no coronavirus related data and the suggestions I made are only based on the consumption "the more business location we can find in a city district, the more dangerous it is :)" so the next stage is to implement the corona data into this model and cluster it again.

Thanks for reading! And enjoy AI