Coursera Capstone

Opening a new Cafe in Kassel, Germany

by: Lars Holbein

May 2020

Introduction

We have a coffee culture in Germany and also in my hometown Kassel. Coffee culture is generally understood to mean that cafes still roast their own coffee, that you can sit down, have a coffee with friends and not take a coffee to go or industrial coffee like Starbucks.

Therefore, there are many cafes or coffee shops in the city. Sitting in a cafe is very popular. There are many types of cafes / coffee shops and everyone has their own style and tries to discover a gap in the market. The largest modern art exhibition in the world takes place in Kassel every 5 years, and a large number of visitors come. Visitors want to decide where are cafes? Which one do I want to visit? For this Problem there are many popular solutions so we don't want to look this problem

Business Problem

The objective of this project is to analyse and select the best location for opening a *new* cafe.

If I want to open one as an investor, where do I do it best where is the density not so high? Where is the best place? To solve this we will use data science methodology an ML like clustering.

Data

- To solve the problem we need a list of neighbourhoods in Kassel. The Wikipage https://de.wikipedia.org/wiki/Kategorie:Stadtteil_von_Kassel (https://de.wikipedia.org/wiki/Kategorie:Stadtteil_von_Kassel) contains this information.
- Latitude and longtime coordinates
- Venue data of cafes

Method

I will extract these data from the page just like in exercise 3. Then generate the coordinates using python geocoder. After that I will use the Foursquare API to get the venue data I am interessted in. I will use data cleaning, ML (clustering), and map Visualisation.

- #### Data Cleaning

First the data of the district of the city of Kassel were extracted from the wiki page (https://de.wikipedia.org/wiki/Kategorie:Stadtteil_von_Kassel (https://de.wikipedia.org/wiki/Kategorie:Stadtteil_von_Kassel)). The coordinates were extracted and merged using geocoder (Geocoder is a simple and consistent geocoding library).

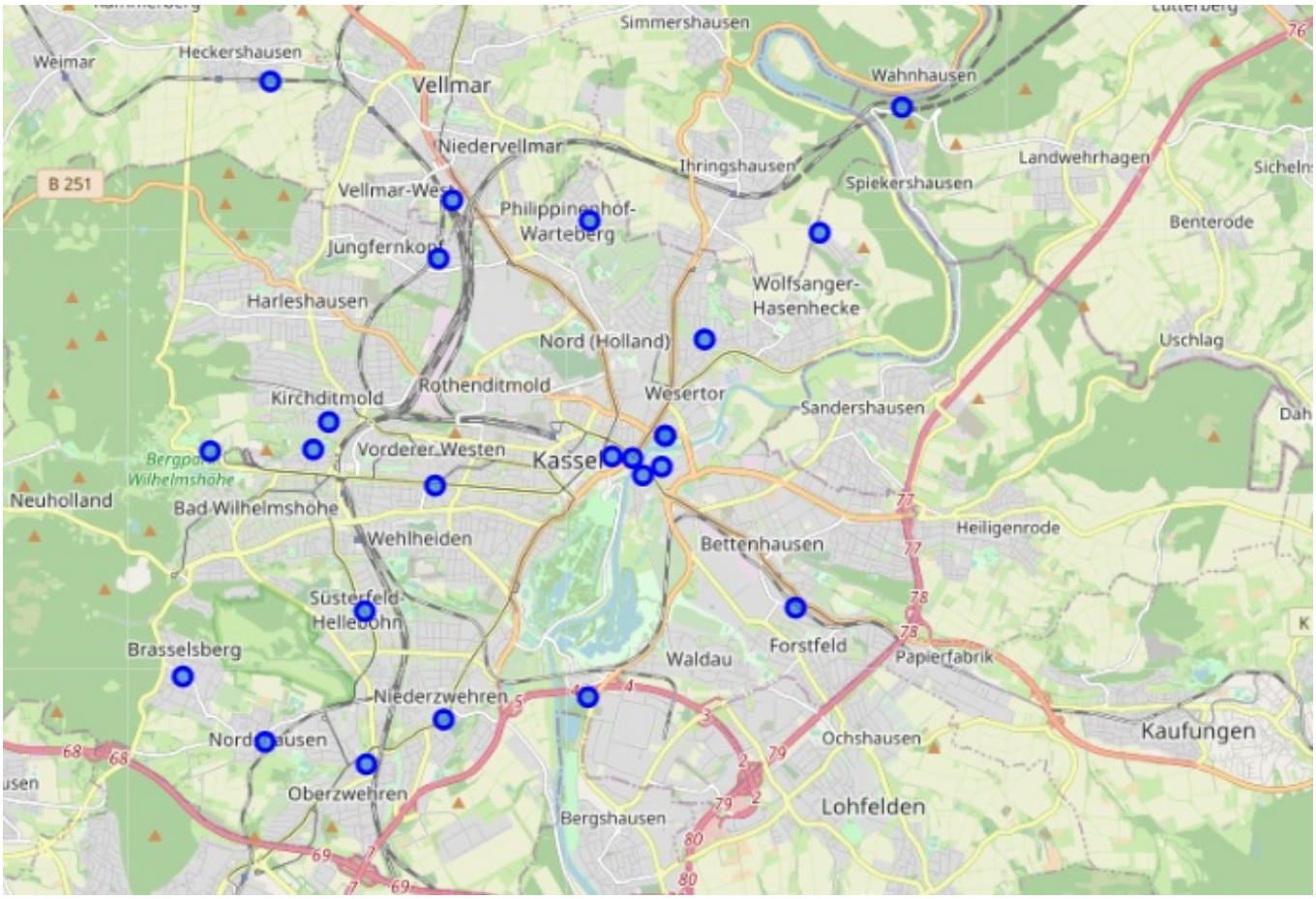
	Neighborhood	Latitude	Longitude
0	Bad Wilhelmshöhe	51.316602	9.420743
1	Bettenhausen (Kassel)	51.313761	9.507324
2	Brasselsberg (Kassel)	51.288650	9.415340
3	Fasanenhof (Kassel)	51.330630	9.519510
4	Forstfeld (Kassel)	51.297133	9.537739
5	Harleshausen	51.362750	9.432810
6	Jungfernkopf	51.340660	9.466560
7	Kirchditmold	51.316970	9.441508
8	Kragenhof	51.359512	9.559258
9	Mitte (Kassel)	51.316027	9.501226
10	Niederzwehren	51.283270	9.467520
11	Nord-Holland (Kassel)	51.318530	9.511830
12	Nordshausen	51.280430	9.431680
13	Oberneustadt	49.242480	8.264840
14	Oberzwehren	51.277600	9.452090
15	Philippinenhof-Warteberg	51.345410	9.496480
16	Rothenditmold	51.348083	9.469283
17	Südstadt (Kassel)	51.318530	9.511830
18	Süsterfeld-Helleböhn	51.296760	9.451780
19	Unterneustadt	51.314700	9.511140
20	Vorderer Westen	51.624627	9.617860
21	Wahlershausen	51.320211	9.444337
22	Waldau (Kassel)	51.286062	9.496285
23	Wehlheiden	51.312324	9.465766
24	Wesertor (Kassel)	51.315808	9.505281
25	Wolfsanger / Hasenhecke	51.343815	9.542803

- #### Feature Selection

Now the proposed venues within the neighborhoods have been considered. The cafe and coffeeshop for Foursqure showed two different types, but in reality they have to be considered as one. This can be safely accepted since I live in the city and know all the suggestions.

	Neighborhoods	African Restaurant	American Restaurant	Art Gallery	Art Museum	Asian Restaurant	Automotive Shop	BBQ Joint	Bagel Shop	Bakery	Bank	Bar	Beer Garden	Big Box Store	Bookstore	Bowling Alley	Breakfast Spot	Brewery	Burger Joint	Bus Stop	Business Service	Café	Castle	Chinese Restaurant	Clothing Store	Cocktail Bar	Coffee Shop
0	Bad Wilhelmshöhe	0.000000	0.027778	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.055556	0.027778	0.000000	0.000000	0.000000	0.027778	0.000000	0.000000	0.027778	0.000000	0.000000	0.000000	0.055556	0.027778	0.027778	0.000000	0.000000	0.027778
1	Bettenhausen (Kassel)	0.010101	0.010101	0.030303	0.040404	0.010101	0.010101	0.00	0.010101	0.020202	0.010101	0.030303	0.010101	0.010101	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.050505	0.000000	0.000000	0.030303	0.040404	0.050505
2	Brasselsberg (Kassel)	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3	Fasanenhof (Kassel)	0.026316	0.026316	0.026316	0.000000	0.000000	0.000000	0.00	0.000000	0.052632	0.000000	0.000000	0.000000	0.026316	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.052632	0.000000	0.000000	0.000000	0.000000	0.026316
4	Forstfeld (Kassel)	0.000000	0.043478	0.000000	0.000000	0.000000	0.043478	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.043478	0.043478	0.000000	0.000000	0.043478	0.000000	0.000000	0.000000
5	Harleshausen	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
6	Jungfernkopf	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.045455	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.045455	0.000000	0.000000	0.000000	0.000000	0.000000
7	Kirchditmold	0.000000	0.023810	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.023810	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.023810	0.023810	0.000000	0.000000	0.071429	0.000000	0.023810	0.000000	0.000000	0.000000
8	Kragenhof	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
9	Mitte (Kassel)	0.010204	0.010204	0.030612	0.040816	0.000000	0.010204	0.00	0.010204	0.020408	0.010204	0.030612	0.010204	0.010204	0.000000	0.000000	0.010204	0.000000	0.000000	0.000000	0.000000	0.051020	0.000000	0.000000	0.030612	0.040816	0.051020
10	Niederzwehren	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.045455	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.045455	0.000000	0.000000
11	Nord-Holland (Kassel)	0.010000	0.010000	0.030000	0.040000	0.000000	0.010000	0.01	0.010000	0.020000	0.010000	0.020000	0.010000	0.010000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.050000	0.000000	0.000000	0.040000	0.030000	0.050000
12	Nordshausen	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
13	Oberneustadt	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.250000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
14	Oberzwehren	0.000000	0.000000	0.000000	0.000000	0.033333	0.000000	0.00	0.000000	0.033333	0.000000	0.000000	0.000000	0.000000	0.033333	0.000000	0.000000	0.033333	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
15	Philippinenhof-Warteberg	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.050000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
16	Rothenditmold	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
17	Südstadt (Kassel)	0.010000	0.010000	0.030000	0.040000	0.000000	0.010000	0.01	0.010000	0.020000	0.010000	0.020000	0.010000	0.010000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.050000	0.000000	0.000000	0.040000	0.030000	0.050000
18	Süsterfeld-Helleböhn	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.057143	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.028571	0.000000	0.000000	0.028571	0.000000	0.028571	0.028571	0.000000	0.000000
19	Unterneustadt	0.010000	0.010000	0.030000	0.040000	0.000000	0.010000	0.01	0.010000	0.020000	0.010000	0.030000	0.010000	0.010000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.040000	0.000000	0.000000	0.040000	0.030000	0.050000
20	Vorderer Westen	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
21	Wahlershausen	0.000000	0.023256	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.023256	0.000000	0.023256	0.000000	0.000000	0.000000	0.000000	0.000000	0.023256	0.023256	0.000000	0.000000	0.069767	0.000000	0.023256	0.000000	0.000000	0.000000
22	Waldau (Kassel)	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.057143	0.000000	0.000000	0.000000	0.028571	0.000000	0.028571	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.028571	0.028571	0.000000	0.000000

- #### Explore Dataset In order to get a better overview, the venues were visualized on the map.



- #### Clustering

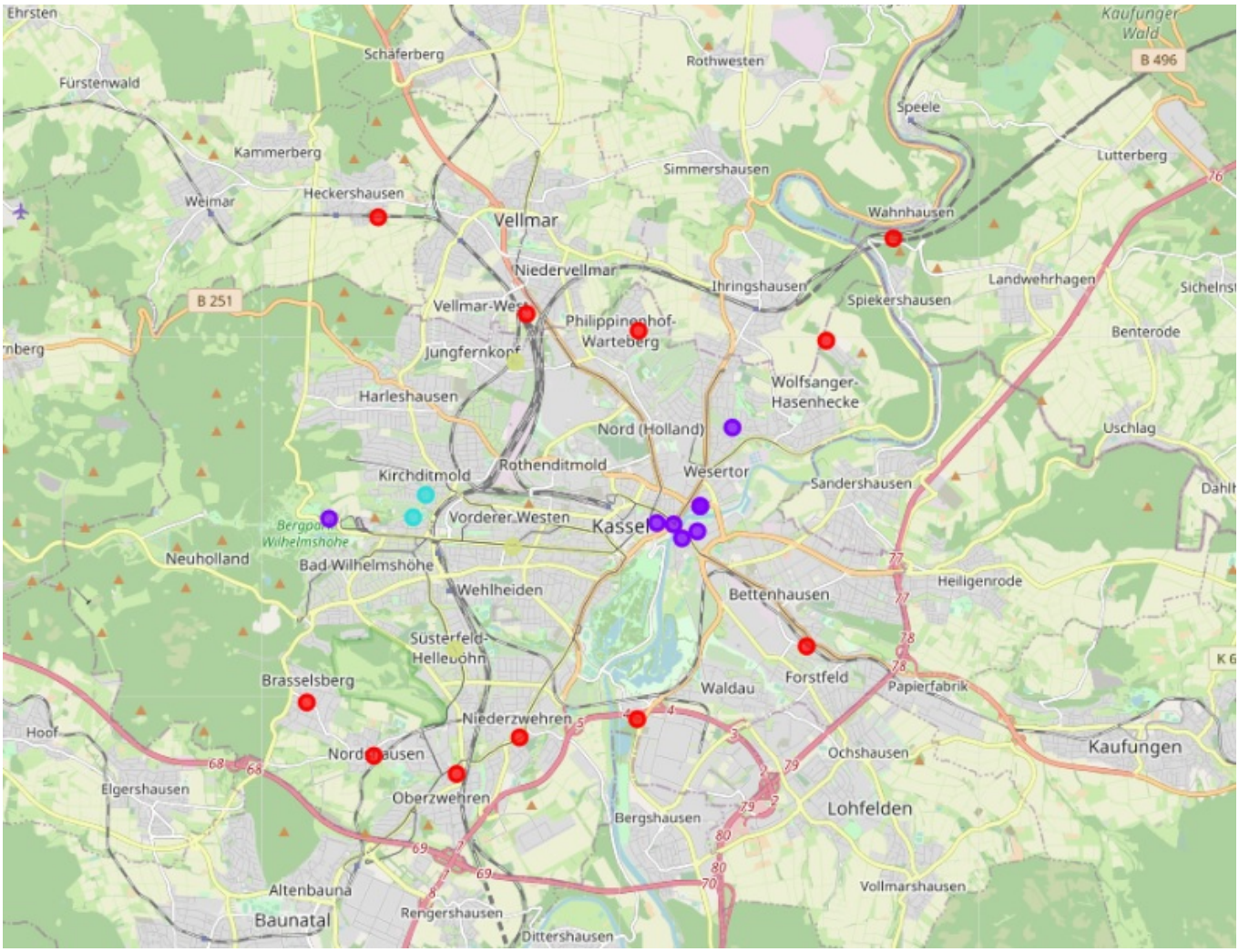
K-means was selected as the cluster algorithm with a k = 3. This seemed to make sense since we have unlabeled data so that we have an unsupervised learning problem. By clustering the districts, patterns were found that found similar districts and were the target.

Results

The Results from the k-mean clustering show that we can catogorize the city into 3 cluster based frequency

- cluster 0 : moderate number of Cafes/Coffeeshop (mint green dot)
- cluster 1 : low number of Cafes/Coffeeshop (red dot)
- cluster 2 : high number of Cafes/Coffeeshop (purple green dot)

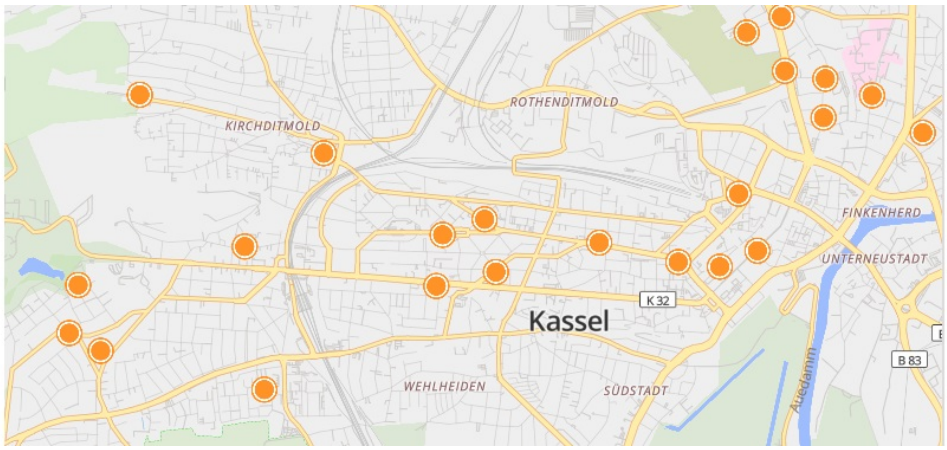
The result of clustering are in the figure below:



Discussion and Conclusion

When looking at the results, one could come to the conclusion that this would be a good place for a new cafe in cluster 1 and the basic avoidance of cluster 2. If only one trusted the data, it would be so when creating the report and trying The identification of the problem for the reopening of a cafe can be made by applying the learned procedure, however, as previously mentioned, it should be used with caution. While Foursquare knows only 20 cafes, a look at a German app already shows a picture of 62 cafes (personal feeling of someone who lives here is even more). The problem here is the data quality of Foursquare. Up to the start of this course I have never heard of foursquare and I do not know any active users of Foursquare. The data quality appears to be good in the USA and Canada, but probably not in Germany. However, since the German portal does not provide an API, it cannot be used.

Insgesamt haben wir 62 Cafes mit 16.311 Bewertungen aus 52 Portalen gefunden



The problem here is the data quality of Foursquare. Up to the start of this course I have never heard of foursquare and I do not know any active users of Foursquare. The data quality appears to be good in the USA and Canada, but probably not in Germany. However, since the German portal does not provide an API, it cannot be used.

In [ ]: