IBM Applied Data Science Capstone Project

The Restaurant Battle of Neighborhoods in Cologne, Germany

By Ayoub Nassiri

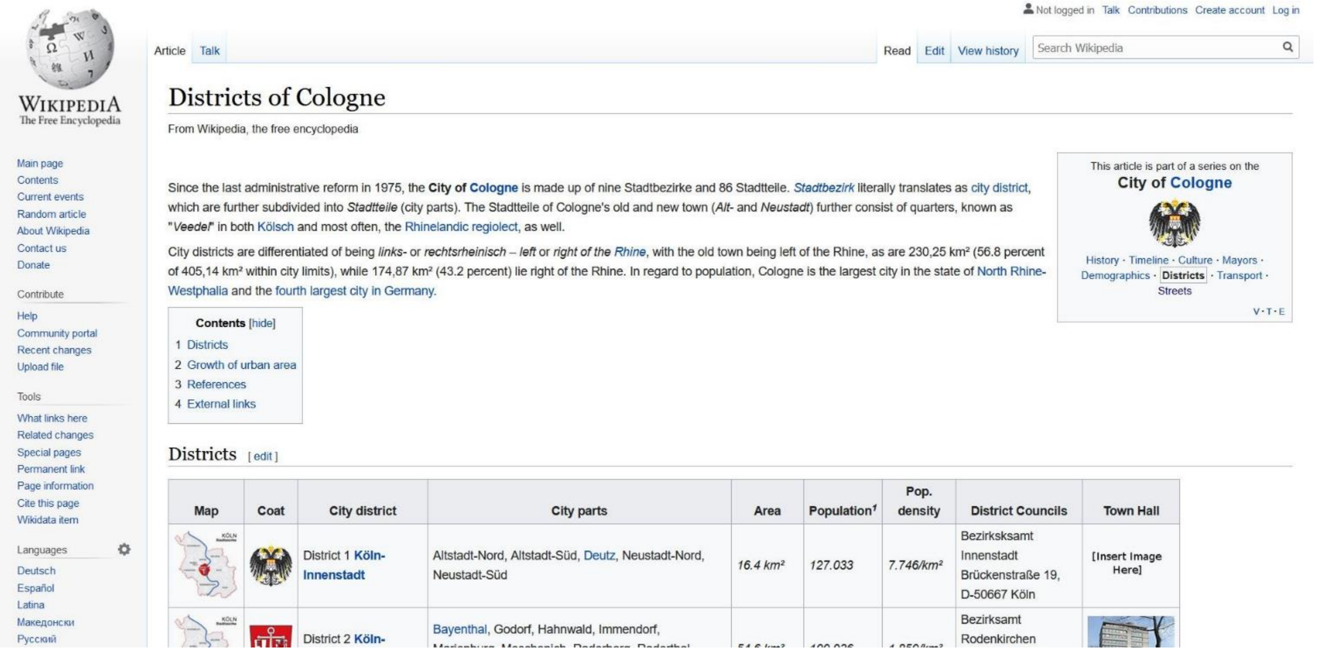
Table of contents

Introduction: Business problem

Cologne, the city the author lives in, attracts a large number of tourists, not least due to its famous cathedral, the trade fairs and conventions, such as the gamescom, and its vibrant party scene. For tourists, finding the right place to eat can be a challenge, though. German dishes include a lot of meat, often pork, which many people do not want to eat for health-related, religious, cultural or moral reasons. This is just one motive for giving tourists a good overview about what to eat where. Thus, the goal I want to reach with this exercise is to give a simple recommendation to tourists in Cologne: in which district of the city will you find a large number or even concentration of which types of restaurants? Where to eat Mediterranean food, where to find German food, where to get fast food? The target audience are foreign tourists.

Data

I will, as requested by the assignment task, use foursquare data about restaurants in Cologne. Foursquare is a US tech company from New York focusing on location data. Their technology and data powers apps such as Apple's Maps, Uber, Twitter and many other household names. Here is an example of a vegetarian restaurant in Cologne on foursquare: https://de.foursquare.com/v/sattgr%C3%BCn/5c33306cc824ae002c2b414c. I will use foursquare data such as the restaurant name, ID, location and category of food (vegetarian, Italian etc.). Also, I will use the overview of districts/city parts of Cologne from Wikipedia: https://en.wikipedia.org/wiki/Districts_of_Cologne.



Methodology

In this section, I will describe the data analysis and how I used the data to yield the results.

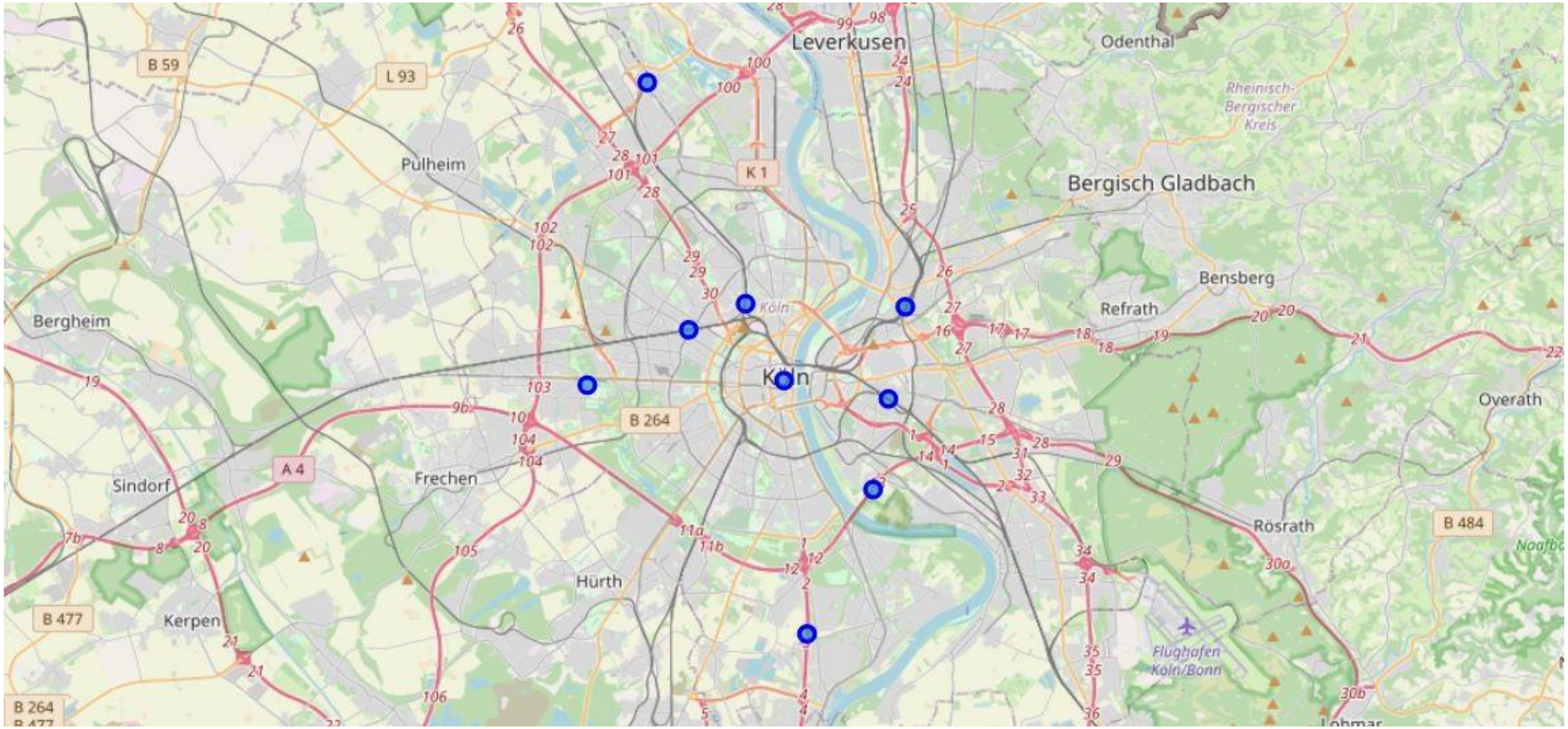
Starting out, I scraped data from Wikipedia to create a dataframe with the city districts of Cologne: https://en.wikipedia.org/wiki/Districts_of_Cologne. For this, I used the pandas read function. I had to clean the resulting data frame in terms of unnecessary information or data that could not be handled in a data frame, such as picture data of the coat of arms of each district. The result is a nice data frame:

	City district	City parts	Area	Population1	Pop. density	District Councils
0	Köln-Innenstadt	Altstadt-Nord, Altstadt-Süd, Deutz, Neustadt-N...	16.4 km²	127.033	7.746/km²	Bezirksksamt Innenstadt Brückenstraße 19, D-50...
1	Köln-Rodenkirchen	Bayenthal, Godorf, Hahnwald, Immendorf, Marien...	54.6 km²	100.936	1.850/km²	Bezirksamt Rodenkirchen Hauptstraße 85, D-5099...
2	Köln-Lindenthal	Braunsfeld, Junkersdorf, Klettenberg, Lindenth...	41.6 km²	137.552	3.308/km²	Bezirksamt Lindenthal Aachener Straße 220, 509...
3	Köln-Ehrenfeld	Bickendorf, Bocklemünd/Mengenich, Ehrenfeld, N...	23.8 km²	103.621	4.348/km²	Bezirksamt Ehrenfeld Venloer Straße 419 – 421,...
4	Köln-Nippes	Bilderstöckchen, Longerich, Mauenheim, Niehl, ...	31.8 km²	110.092	3.462/km²	Bezirksamt NippesNeusser Straße 450,D-50733 Köln
5	Köln-Chorweiler	Blumenberg, Chorweiler, Esch/Auweiler, Fühling...	67.2 km²	80.870	1.204/km²	Bezirksamt Chorweiler Pariser Platz 1, D-50765...
6	Köln-Porz	Eil, Elsdorf, Ensen, Finkenberg, Gremberghoven...	78.8 km²	106.520	1.352/km²	Bezirksamt PorzFriedrich-Ebert-Ufer 64–70, D-5...
7	Köln-Kalk	Brück, Höhenberg, Humboldt/Gremberg, Kalk, Mer...	38.2 km²	108.330	2.841/km²	Bezirksamt KalkKalker Hauptstraße 247–273,D-51...
8	Köln-Mülheim	Buchforst, Buchheim, Dellbrück, Dünnwald, Flit...	52.2 km²	144.374	2.764/km²	Bezirksamt Mülheim Wiener Platz 2a,D-51065 Köln

Then, I enabled geopy functions by installing the conda-forge geopy package. I used the nominatim function to add geospatial data to the data frame, that is the latitude and the longitude seen on the right side of the following table.

	City district	City parts	Area	Population1	Pop. density	District Councils	Latitude	Longitude
0	Köln-Innenstadt	Altstadt-Nord, Altstadt-Süd, Deutz, Neustadt-N...	16.4 km²	127.033	7.746/km²	Bezirksksamt Innenstadt Brückenstraße 19, D-50...	50.937328	6.959234
1	Köln-Rodenkirchen	Bayenthal, Godorf, Hahnwald, Immendorf, Marien...	54.6 km²	100.936	1.850/km²	Bezirksamt Rodenkirchen Hauptstraße 85, D-5099...	50.865622	6.969718
2	Köln-Lindenthal	Braunsfeld, Junkersdorf, Klettenberg, Lindenth...	41.6 km²	137.552	3.308/km²	Bezirksamt Lindenthal Aachener Straße 220, 509...	50.935935	6.871246
3	Köln-Ehrenfeld	Bickendorf, Bocklemünd/Mengenich, Ehrenfeld, N...	23.8 km²	103.621	4.348/km²	Bezirksamt Ehrenfeld Venloer Straße 419 – 421,...	50.951502	6.916529
4	Köln-Nippes	Bilderstöckchen, Longerich, Mauenheim, Niehl, ...	31.8 km²	110.092	3.462/km²	Bezirksamt NippesNeusser Straße 450,D-50733 Köln	50.958994	6.941777
5	Köln-Chorweiler	Blumenberg, Chorweiler, Esch/Auweiler, Fühling...	67.2 km²	80.870	1.204/km²	Bezirksamt Chorweiler Pariser Platz 1, D-50765...	51.021167	6.898034
6	Köln-Porz	Eil, Elsdorf, Ensen, Finkenberg, Gremberghoven...	78.8 km²	106.520	1.352/km²	Bezirksamt PorzFriedrich-Ebert-Ufer 64–70, D-5...	50.906705	6.999129
7	Köln-Kalk	Brück, Höhenberg, Humboldt/Gremberg, Kalk, Mer...	38.2 km²	108.330	2.841/km²	Bezirksamt KalkKalker Hauptstraße 247–273,D-51...	50.931923	7.005806
8	Köln-Mülheim	Buchforst, Buchheim, Dellbrück, Dünnwald, Flit...	52.2 km²	144.374	2.764/km²	Bezirksamt Mülheim Wiener Platz 2a,D-51065 Köln	50.958147	7.013526

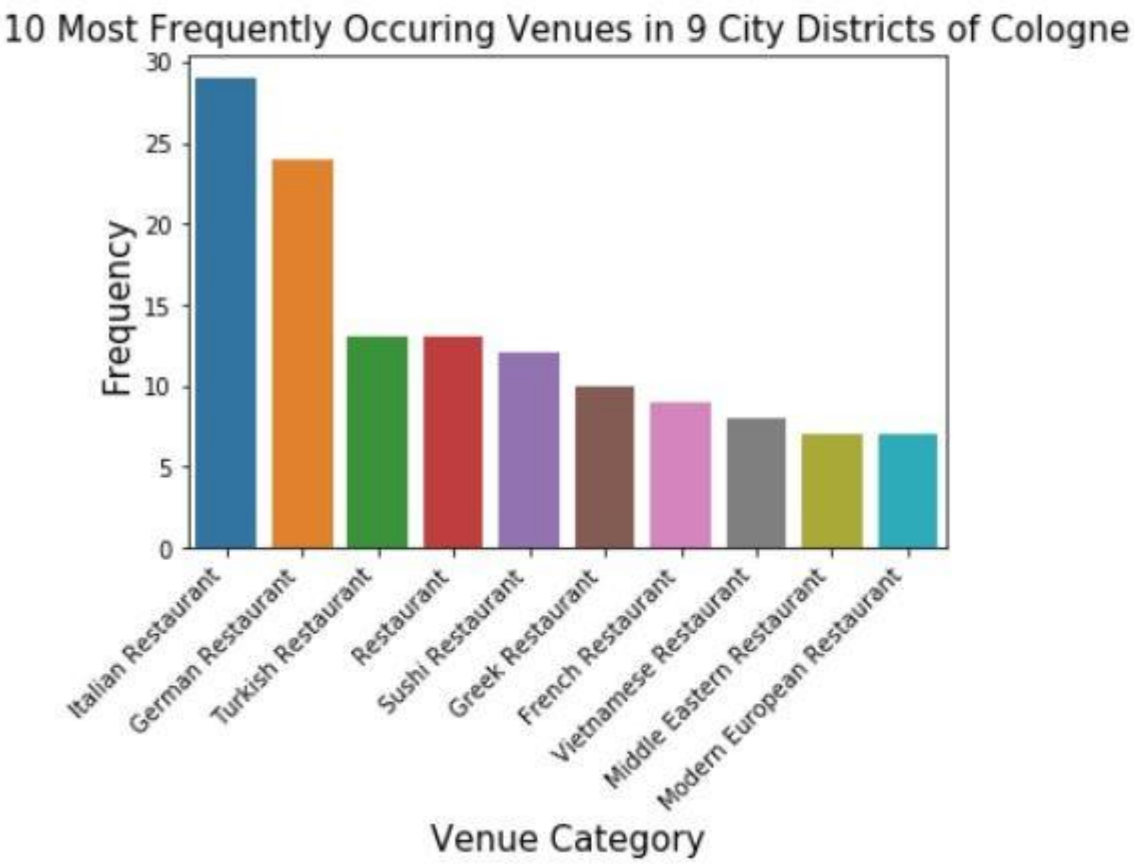
Using the folium package and my data frame, I then created a map with the nine city districs on it.



Now, foursquare data comes into play. I first did a view try-outs for the city district "Innenstadt", which I know pretty well, to see if the venues retrieved from foursquare seem reasonable and correct. That was the case.

Then, retrieved the foursquare data for all venues on foursquare with a distance of less than 3000 meters from each center of each city district, as indicated as blue dots in the map above. The result was a list of 757 venues all over Cologne city. Out of these 757 venues, 184 where restaurants. These 184 restaurants come from 35 unique restaurant categories, such as Italian, Vietnamese or German.

I plotted a bar chart with the frequency of the 10 most frequently occuring restaurants in the whole city, using seaborn/matplotlib packages. We can see that Italian, German and Turkish restaurants are the most frequently occuring restaurants in Cologne, which seems pretty reasonable, as Germany has a relatively high proportion of people with Italian and Turkish roots, and these cuisines being excellent and highly appreciated by large parts of the population - think about pizza, pasta or kebap!



To find clusters of restaurant types in the different city districts, I first transformed the data frame with the restaurant venues, associated to city districts, by one-hot encoding (0/1), as seen in the picture below.

	Neighborhood	American Restaurant	Asian Restaurant	Austrian Restaurant	Chinese Restaurant	Comfort Food Restaurant	Doner Restaurant	Eastern European Restaurant	Falafel Restaurant	Fast Food Restaurant	French Restaurant	German Restaurant	Greek Restaurant	Indian Restaurant	Italian Restaurant	Japanese Restaurant	Kebab Restaurant	Kurdish Restaurant	A Re
1	Köln-Innenstadt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	Köln-Innenstadt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	Köln-Innenstadt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	Köln-Innenstadt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5	Köln-Innenstadt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Next, I used grouping to show the frequency of each category of restaurants in each city district.

	Neighborhood	American Restaurant	Asian Restaurant	Austrian Restaurant	Chinese Restaurant	Comfort Food Restaurant	Doner Restaurant	Eastern European Restaurant	Falafel Restaurant	Fast Food Restaurant	French Restaurant	German Restaurant	Greek Restaurant	Indian Restaurant	Italian Restaurant	Japanese Restaurant	Kebab Restaurant	Kurdish Restaurant
0	Köln-Chorweiler	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.333333	0.000000	0.000000	0.000000	0.000000	0.333333	0.000000	0.000000	0.000000
1	Köln-Ehrenfeld	0.000000	0.000000	0.000000	0.038462	0.000000	0.000000	0.000000	0.000000	0.000000	0.038462	0.076923	0.038462	0.000000	0.153846	0.038462	0.038462	0.000000
2	Köln-Innenstadt	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.041667	0.000000	0.083333	0.083333	0.000000	0.000000	0.291667	0.000000	0.000000	0.000000
3	Köln-Kalk	0.000000	0.041667	0.000000	0.000000	0.041667	0.041667	0.041667	0.000000	0.000000	0.000000	0.041667	0.083333	0.041667	0.083333	0.000000	0.000000	0.041667
4	Köln-Lindenthal	0.041667	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.041667	0.333333	0.083333	0.041667	0.083333	0.041667	0.000000	0.000000
5	Köln-Mülheim	0.000000	0.095238	0.000000	0.000000	0.047619	0.000000	0.000000	0.000000	0.000000	0.000000	0.047619	0.047619	0.000000	0.095238	0.000000	0.000000	0.000000
6	Köln-Nippes	0.000000	0.000000	0.064516	0.000000	0.000000	0.032258	0.000000	0.000000	0.000000	0.096774	0.064516	0.064516	0.000000	0.193548	0.000000	0.032258	0.000000
7	Köln-Porz	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.035714	0.071429	0.250000	0.071429	0.000000	0.178571	0.000000	0.000000	0.035714
8	Köln-Rodenkirchen	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.333333	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

I used this information to create a data frame in which you can see the most common restaurant venue types for each city district.

	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	Köln-Chorweiler	Fast Food Restaurant	Italian Restaurant	Sushi Restaurant	Vietnamese Restaurant	Japanese Restaurant	Indian Restaurant	Greek Restaurant	German Restaurant	French Restaurant	Falafel Restaurant
1	Köln-Ehrenfeld	Tapas Restaurant	Italian Restaurant	Restaurant	German Restaurant	Sushi Restaurant	Vietnamese Restaurant	Modern European Restaurant	Chinese Restaurant	French Restaurant	Greek Restaurant
2	Köln-Innenstadt	Italian Restaurant	Sushi Restaurant	Vietnamese Restaurant	German Restaurant	French Restaurant	Middle Eastern Restaurant	Modern European Restaurant	Restaurant	Schnitzel Restaurant	Falafel Restaurant
3	Köln-Kalk	Greek Restaurant	Turkish Restaurant	Italian Restaurant	Middle Eastern Restaurant	Restaurant	German Restaurant	Indian Restaurant	Kurdish Restaurant	Vegetarian / Vegan Restaurant	Mediterranean Restaurant
4	Köln-Lindenthal	German Restaurant	Sushi Restaurant	Italian Restaurant	Greek Restaurant	American Restaurant	Indian Restaurant	French Restaurant	Mexican Restaurant	Modern European Restaurant	Restaurant
5	Köln-Mülheim	Turkish Restaurant	Italian Restaurant	Asian Restaurant	Mediterranean Restaurant	German Restaurant	Vegetarian / Vegan Restaurant	Greek Restaurant	Middle Eastern Restaurant	Vietnamese Restaurant	Seafood Restaurant
6	Köln-Nippes	Italian Restaurant	French Restaurant	Vietnamese Restaurant	Austrian Restaurant	Modern European Restaurant	Greek Restaurant	Restaurant	Sushi Restaurant	German Restaurant	Spanish Restaurant
7	Köln-Porz	German Restaurant	Italian Restaurant	Restaurant	Greek Restaurant	Thai Restaurant	Seafood Restaurant	French Restaurant	Fast Food Restaurant	Turkish Restaurant	Kurdish Restaurant
8	Köln-Rodenkirchen	German Restaurant	Restaurant	Scandinavian Restaurant	Vietnamese Restaurant	Kebab Restaurant	Italian Restaurant	Indian Restaurant	Greek Restaurant	French Restaurant	Fast Food Restaurant

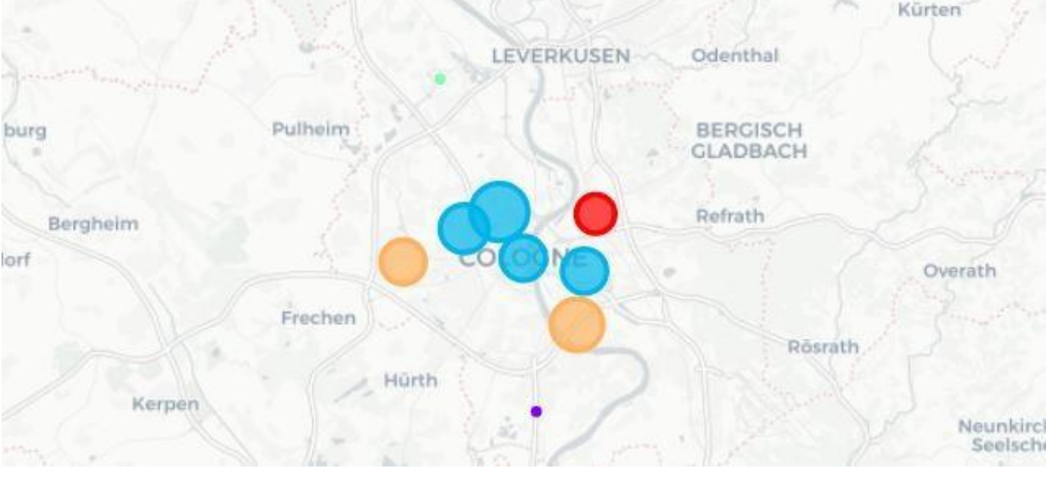
Now, with all this data, I could finally run an unsupervised machine learning algorithm, more specifically, a k-means clustering algorithm from the scikit-learn package. One could use the ellbow method to systematically define the k value, but I simply chose k to be 5, having been inspired by one of the coursera courses to do so.

Results

And here already comes the result:

	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 Mc
0	3	Köln-Chorweiler	Fast Food Restaurant	Italian Restaurant	Sushi Restaurant	Vietnamese Restaurant	Japanese Restaurant	Indian Restaurant	Gree
1	2	Köln-Ehrenfeld	Tapas Restaurant	Italian Restaurant	Restaurant	German Restaurant	Sushi Restaurant	Vietnamese Restaurant	Mode
2	2	Köln-Innenstadt	Italian Restaurant	Sushi Restaurant	Vietnamese Restaurant	German Restaurant	French Restaurant	Middle Eastern Restaurant	Mode
3	2	Köln-Kalk	Greek Restaurant	Turkish Restaurant	Italian Restaurant	Middle Eastern Restaurant	Restaurant	German Restaurant	India
4	4	Köln-Lindenthal	German Restaurant	Sushi Restaurant	Italian Restaurant	Greek Restaurant	American Restaurant	Indian Restaurant	Frenc
5	0	Köln-Mülheim	Turkish Restaurant	Italian Restaurant	Asian Restaurant	Mediterranean Restaurant	German Restaurant	Vegetarian / Vegan Restaurant	Gree
6	2	Köln-Nippes	Italian Restaurant	French Restaurant	Vietnamese Restaurant	Austrian Restaurant	Modern European Restaurant	Greek Restaurant	
7	4	Köln-Porz	German Restaurant	Italian Restaurant	Restaurant	Greek Restaurant	Thai Restaurant	Seafood Restaurant	Frenc
8	1	Köln-Rodenkirchen	German Restaurant	Restaurant	Scandinavian Restaurant	Vietnamese Restaurant	Kebab Restaurant	Italian Restaurant	India

What we see in the table are the city districts and their most common venues, and they now have been assigned five different cluster labels from 0 to 4. We can now use the cluster labels to show the city districts marked with a cluster-specific color on a map, again using folium:



You will see nine bubbles for the nine city districts, with five different colors for the five different clusters. If you have trouble counting to five here, look for a small green dot on the upper part of the picture and a small purple dot on the lower part of the picture.

Now, what is the final result of this exercise? We now can show five clusters of restaurant type concentrations for the city of Cologne, which I named according to the restaurant concentration the data shows.

Cluster 1 - the Turkish Food Cluster (Mülheim)

	City parts	District Councils	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th M Comi Ve
8	Buchforst, Buchheim, Dellbrück, Dünnwald, Flit...	Bezirksamt Mülheim Wiener Platz 2a,D-51065 Köln	50.958147	7.013526	0	Turkish Restaurant	Italian Restaurant	Asian Restaurant	Mediterranean Restaurant	German Restaurant	Vegetar Ve Restau

Cluster 2 - the Northern European Food Cluster (Rodenkirchen)

	City parts	District Councils	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th M Comm Ver
1	Bayenthal, Godorf, Hahnwald, Immendorf, Marien...	Bezirksamt Rodenkirchen Hauptstraße 85, D-5099...	50.865622	6.969718	1	German Restaurant	Restaurant	Scandinavian Restaurant	Vietnamese Restaurant	Kebab Restaurant	Ita Restaur

Cluster 3 - the Mediterranean Food Cluster (Innenstadt, Ehrenfeld, Nippes, Kalk)

	City parts	District Councils	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th M Comn Ver
0	Altstadt-Nord, Altstadt-Süd, Deutz, Neustadt-N...	Bezirksksamt Innenstadt Brückenstraße 19, D-50...	50.937328	6.959234	2	Italian Restaurant	Sushi Restaurant	Vietnamese Restaurant	German Restaurant	French Restaurant	Mic East Restau
3	Bickendorf, Bocklemünd/Mengenich, Ehrenfeld, N...	Bezirksamt Ehrenfeld Venloer Straße 419 – 421,...	50.951502	6.916529	2	Tapas Restaurant	Italian Restaurant	Restaurant	German Restaurant	Sushi Restaurant	Vietnam Restau
4	Bilderstöckchen, Longerich, Mauenheim, Niehl, ...	Bezirksamt NippesNeusser Straße 450,D-50733 Köln	50.958994	6.941777	2	Italian Restaurant	French Restaurant	Vietnamese Restaurant	Austrian Restaurant	Modern European Restaurant	Gr Restau
7	Brück, Höhenberg, Humboldt/Gremberg, Kalk, Mer...	Bezirksamt KalkKalker Hauptstraße 247–273,D-51...	50.931923	7.005806	2	Greek Restaurant	Turkish Restaurant	Italian Restaurant	Middle Eastern Restaurant	Restaurant	Gern Restau

Cluster 4 - the Fast Food Cluster (Chorweiler)

	City parts	District Councils	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Mo Comm Ven
5	Blumenberg, Chorweiler, Esch/Auweiler, Fühling...	Bezirksamt Chorweiler Pariser Platz 1, D-50765...	51.021167	6.898034	3	Fast Food Restaurant	Italian Restaurant	Sushi Restaurant	Vietnamese Restaurant	Japanese Restaurant	Ind Restaur

Cluster 5 - the German/Diverse Food Cluster (Lindenthal, Porz)

	City parts	District Councils	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Mo Comm Ven
2	Braunsfeld, Junkersdorf, Klettenberg, Lindenth...	Bezirksamt Lindenthal Aachener Straße 220, 509...	50.935935	6.871246	4	German Restaurant	Sushi Restaurant	Italian Restaurant	Greek Restaurant	American Restaurant	Indi Restaura
6	Eil, Elsdorf, Ensen, Finkenberg, Gremberghoven...	Bezirksamt PorzFriedrich-Ebert-Ufer 64–70, D-5...	50.906705	6.999129	4	German Restaurant	Italian Restaurant	Restaurant	Greek Restaurant	Thai Restaurant	Seafo Restaura

Interestingly, it is really possible to define clusters of certain cuisines in Cologne city. People living in Cologne will probably agree that these clusters sound pretty reasonable and are not too far away from what you would have expected.

Discussion

If I reflect the work necessary to create these results, what comes to my mind is that for typical ways of scraping, cleaning, handling, transforming and visualizing data, all the tools are simply there. We just have to get to know the available open source packages and learn how to use them. What I find fantastic is that nearly all of them are free of charge. Also, a simple notebook computer is enough: in my case, I used a ThinkPad L470, more than three years old. All the rest is concentrated, creative, interesting, sometimes hard work and searching for hints, tips, examples, explanations etc. in the web. With these tools, many exciting data science use cases can be created, for all kinds of useful purposes.

Conclusion

We achieved the goal presented at the outset of this blogpost: tourists can see in the results which city districts best match their food desires. This is just one example of fantastic data science uses cases one can realize applying technology which is available for free today! What a time to be alive.