Capstone Project - The Battle of Neighborhoods

by Stephan Bartelheim

1 Introduction

Düsseldorf is a German city with the highest concentration of Japanese citizens outside Japan. Apart from a big number of subsidiaries of big Japanese companies and banks, there is also a plethora of Japanese shops and restaurants. Our client, a recent arrival from Japan, wishes to open another Japanese restaurant in the city. Since he is unfamiliar with the place, he needs advice on where to open and what kind of food to offer. We, therefore, try to identify areas in which Japanese restaurants tend to thrive and then pick one where the competition is particularly weak. Lastly we will choose a category of Japanese food (e.g. Ramen, Sushi, general) that is generally popular but again faces weak competition in our selected area. In the end the client will be left with a very specific recommendation.

2 Data

The area data can be accessed under https://www.dasoertliche.de/Themen/Postleitzahlen/D%C3%BCsseldorf.html. It is from the German phone register provided by the *Deutsche Telekom AG* and is free to use. Unlike in the previous exercises we will divide the city by postal codes and not by neighborhoods, since that is how the data is structured. We will use the *geocoder package* to add coordinates and the FourSquare API *explore* call to retrieve data about the composition of the areas. This will give us a list of venues listed on FourSquare. Apart from businesses this can for example be tourist sites or sites for access to public transport. Lastly, the data about the competing venues in the chosen area is retrieved from the FourSqare API to see how those businesses are rated by customers.

3 Methodology

We scrape the postal code area data from *DasOertliche,* remove unnecessary columns, add coordinates with geocoder package and the arcgis method and retrieve venues using the FourSquare API explore call in a 500 metre radius. Looking at the number of venues in the different areas (Table next page) we see that there is many areas with very few venues. We have in other words very few information about the area. This also makes clustering those areas very high variance. We therefore discard areas with fewer than 15 listed venues. This reduces the number of potential areas from 38 to 14 and the total number of venues to 850. Those 850 venues, however, come from 167 categories. Using relative frequencies of all those categories would again make the clustering very high variance. To prevent this we

	Neighborhood Latitude	Neighborhood Longitude	Venue_ID	Venue	Venue Latitude	Venue Longitude	Venue Category
Postal Code							
40210	100	100	100	100	100	100	100
40211	38	38	38	38	38	38	38
40212	100	100	100	100	100	100	100
40213	100	100	100	100	100	100	100
40215	64	64	64	64	64	64	64
40217	70	70	70	70	70	70	70
40219	79	79	79	79	79	79	79
40221	2	2	2	2	2	2	2
40223	21	21	21	21	21	21	21
40225	7	7	7	7	7	7	7
40227	26	26	26	26	26	26	26
40229	10	10	10	10	10	10	10
40231	4	4	4	4	4	4	4
40233	26	26	26	26	26	26	26
40235	8	8	8	8	8	8	8
40239	15	15	15	15	15	15	15
40468	4	4	4	4	4	4	4
40470	6	6	6	6	6	6	6
40472	2	2	2	2	2	2	2
40474	4	4	4	4	4	4	4
40476	48	48	48	48	48	48	48
40477	58	58	58	58	58	58	58
40479	53	53	53	53	53	53	53
40489	2	2	2	2	2	2	2
40545	31	31	31	31	31	31	31
40547	12	12	12	12	12	12	12
40549	7	7	7	7	7	7	7
40589	2	2	2	2	2	2	2
40591	11	11	11	11	11	11	11
40593	5	5	5	5	5	5	5
40595	7	7	7	7	7	7	7
40597	36	36	36	36	36	36	36
40599	5	5	5	5	5	5	5
40625	14	14	14	14	14	14	14
40627	4	4	4	4	4	4	4
40629	5	5	5	5	5	5	5

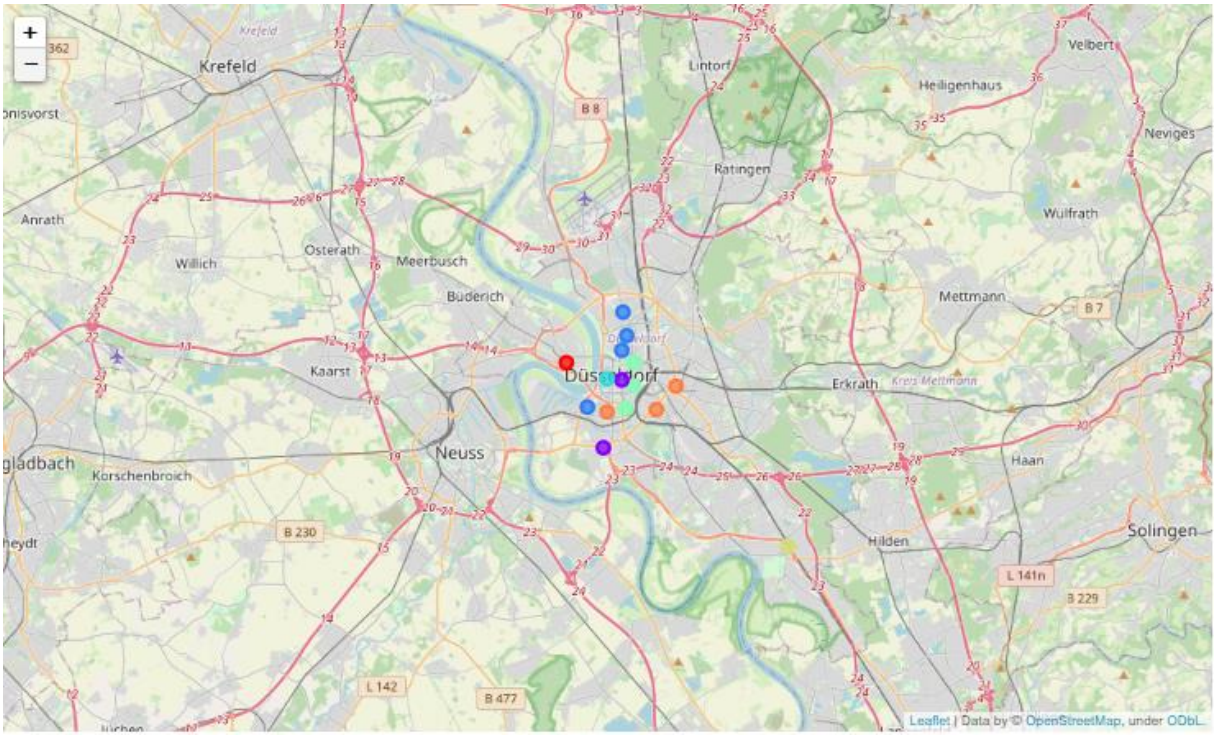
Count of Venues by Postal Code

could again use only the 10 most frequent categories per area and effectively discard the rest or we can group categories to reduce the number to something manageable. Here we choose the second approach. The grouping approach is shown below.

```
#grouping of categories for simplified categories
dus_venues_red['Category simple'].replace(regex=['^Ramen.*','^Sushi.*','^Japanese.*','^Soba.*'] , value='Japanese',inplace=True)
dus_venues_red['Category simple'].replace(regex=['.*[Ss]hop.*','.*[Ss]tore.*','.*Boutique.*'] , value='Shopping',inplace=True)
dus_venues_red['Category simple'].replace(regex=['.*Restaurant.*','^Pizza.*','^Burger.*','^Bistro.*','^Steak.*','^Trattoria.
*','^BBQ.*','^Deli.*','^Sandwich.*','^Breakfast.*','^Soup.*'], value='Restaurant other',inplace=True)
dus_venues_red['Category simple'].replace(regex=['.*Bar.*','.*[Pp]ub.*','^Taverna.*','^Nightclub','^Brewery.*','^Rock.*'], value
='Drinking Place',inplace=True)
dus_venues_red['Category simple'].replace(regex=['.*[Gg]ym.*','.*[Ss]occer.*','.*[Ss]port.*','.*Yoga.*','.*Hockey.*'], value='Sp
orts Venue',inplace=True)
dus_venues_red['Category simple'].replace(regex=['.*[Mm]arket.*','.*[Gg]rocer.*'], value='Groceries',inplace=True)
dus_venues_red['Category simple'].replace(regex=['.*[Mm]useum.*','.*[Tt]heater.*','.*[Gg]allery.*','.*Site.*','^Opera.*'], value
='Culture',inplace=True)
dus_venues_red['Category simple'].replace(regex=['.*[Ss]top.*','.*[Ss]tation.*','[Pp]latform.*'], value='Public Transport',inpla
ce=True)
dus_venues_red['Category simple'].replace(regex=['.*Hostel.*'], value='Hotel',inplace=True)
dus_venues_red['Category simple'].replace(regex=['.*[Pp]ark.*','.*Playground.*','.*[Pp]laza.*','^Fountain.*'], value='Recreatio
n',inplace=True)
dus_venues_red['Category simple'].replace(regex=['.*[Aa]gency.*','.*[Ss]ervice.*'], value='Business Services',inplace=True)
len(dus_venues_red['Category simple'].unique())
```

Grouping of Venue Categories

After the grouping categories with less than 3 venues are discarded and we are left 825 venues. For these we calculate relative frequencies for all postal code areas and use these to knn cluster the areas. We do not use the “Japanese” category, however, because we want to identify neighborhoods that support many Japanese restaurants by their other characteristics. We choose seven clusters since this gives relatively balanced clusters and a sound geographical distribution.



The light mint cluster center-east features on average the highest share of Japanese restaurants among all venues (~10%). So there seems to be characteristics favorable to Japanese restaurants. On the other side it means that our client would face stiff competition

here. The area furthest to south (postal code 40215), however, only has one single Japanese restaurant, serving unspecified Japanese food, with a rating of 7.6.

	Postal Code	Neighborhood Latitude	Neighborhood Longitude	Venue_ID	Venue	Venue Latitude	Venue Longitude	Venue Category	Category simple
1	40210	51.221500	6.789191	4b3b8db9f964a520b27525e3	Kushi Tei of Tokyo	51.223275	6.789558	Japanese Restaurant	Japanese
2	40210	51.221500	6.789191	4b448154f964a520bef525e3	Kagaya	51.221320	6.788232	Japanese Restaurant	Japanese
3	40210	51.221500	6.789191	5053696ce4b08e1d3c985b79	Nagomi	51.221913	6.786502	Japanese Restaurant	Japanese
4	40210	51.221500	6.789191	4b3be8f6f964a520207e25e3	Takumi	51.223429	6.788531	Ramen Restaurant	Japanese
21	40210	51.221500	6.789191	4b7e729cf964a52072ed2fe3	Hyuga	51.224525	6.789297	Sushi Restaurant	Japanese
24	40210	51.221500	6.789191	4f115acee4b09e81d8909f8a	Waraku	51.223664	6.787536	Japanese Restaurant	Japanese
26	40210	51.221500	6.789191	53429155498e8dd5982f0e8e	Takezo Ramen Bar	51.222617	6.790507	Ramen Restaurant	Japanese
27	40210	51.221500	6.789191	5b0c5e3a31ac6c002c6c3868	Tonkatsu Gonta	51.223490	6.788785	Japanese Restaurant	Japanese
32	40210	51.221500	6.789191	4bb64cb06edc76b092b7301c	Naniwa	51.224915	6.788172	Ramen Restaurant	Japanese
34	40210	51.221500	6.789191	59da041146e1b64f7c9cbe9f	Takumi 3rd Tori & Veggie	51.224532	6.788735	Ramen Restaurant	Japanese
36	40210	51.221500	6.789191	4b65c6a0f964a520f8fe2ae3	Yabase	51.224732	6.788633	Japanese Restaurant	Japanese
39	40210	51.221500	6.789191	4bc4a7d7f8219c74eea8b710	Naniwa Sushi & More	51.224845	6.788217	Sushi Restaurant	Japanese
42	40210	51.221500	6.789191	50cb1954e4b0e9d62c78c910	Takumi Tonkotsu & Gyoza (麺処 匠 二代目) (Takumi Ton...	51.225223	6.788210	Japanese Restaurant	Japanese
46	40210	51.221500	6.789191	4bf178b46bfe0f47bdc6d838	Soba-An	51.224914	6.787977	Soba Restaurant	Japanese
48	40210	51.221500	6.789191	562e3587498e62a645809d9e	Maruyasu	51.224577	6.785816	Japanese Restaurant	Japanese
57	40210	51.221500	6.789191	558ef0bd498e512b2db575c7	Yaki The Emon 焼左衛門	51.224743	6.788877	Japanese Restaurant	Japanese
64	40210	51.221500	6.789191	4b4e2b6af964a520e0e326e3	Nagaya	51.225381	6.786096	Japanese Restaurant	Japanese
68	40210	51.221500	6.789191	58515033f7f8dd196fbe8f2e	eat Tokyo	51.223135	6.790070	Ramen Restaurant	Japanese
69	40210	51.221500	6.789191	5810df5538faaa7ba8820c50	Yoshi	51.224217	6.785603	Japanese Restaurant	Japanese
70	40210	51.221500	6.789191	4c0d34d8c700c9b6950ea2dd	Fujiyama	51.219727	6.784333	Sushi Restaurant	Japanese
74	40210	51.221500	6.789191	4cf014b4ed62721ed93864fd	Okinii	51.223453	6.788300	Japanese Restaurant	Japanese
85	40210	51.221500	6.789191	5bbe4960bed483002c64a0f1	Tokyo Ramen	51.224120	6.787225	Ramen Restaurant	Japanese
104	40211	51.229510	6.789159	515080378aca1877ebad8170	sumi.	51.231848	6.792566	Japanese Restaurant	Japanese
118	40211	51.229510	6.789159	50cb1954e4b0e9d62c78c910	Takumi Tonkotsu & Gyoza (麺処 匠 二代目) (Takumi Ton...	51.225223	6.788210	Japanese Restaurant	Japanese
367	40215	51.213835	6.784225	503541ffe4b0bd14213316db	Tokyo Lounge	51.217971	6.782266	Japanese Restaurant	Japanese

Japanese Restaurants in the light mint Cluster

We further see that a high share of the Japanese restaurants in the other areas are Ramen restaurants.

5 Results

We have identified a cluster of three similarly structured adjacent postcode areas in which Japanese restaurants appear widely popular with an average density of Japanese restaurants of almost 10% of all listed venues. Within this cluster, however, we made out one cluster with only one Japanese restaurant, offering unspecified Japanese cuisine, with an average rating 7.6. This seems to be a very promising area to open another restaurant. We further discovered that Ramen restaurants are the next common after the generalists. We would therefore recommend our client to open a Ramen joint to further avoid competition.

6 Discussion

While the results of our analysis might be a good starting point there is several things to consider. Firstly, we had to exclude several neighborhoods from our analysis for a lack of data. In some areas we received information from the FourSquare API for a mere two venues. This might be because there is really few venues in that area or because the venues are missing on FourSquare or just because the 500 metre radius is too small. This leads to the second problem that the area in a 500 metre radius arount the centre of an area might not capture the area very well, either because it is smaller, bigger or just not very circular. Here a different method to retrieve venues might be preferable. Lastly, we do not know if the existing venue structure really predicts the success of a new business very well. Demographic data for example might be more useful.

7 Conclusion

Our recommendation to open a Ramen restaurant in the postal code area 40215 looks sound considering the underlying data. However, with more data sources and some refinements in the analysis we could gain more confidence in our advice.