Coursera
IBM Data Science Specialization

Final Capstone Project Report

Clustering Neighborhoods in Munich

Submitted by:	Patrick Brus
Submission date:	10.04.2020
Course of studys:	IBM Data Science Professional Certificate

Final Capstone Project Report

1 Introduction

1.1 Bussiness Problem

An individual wants to move to munich or wants to find a new apartment in munich. Before he decides for an apartment, he first wants to get some information about the districts in munich and how similar they are. He wants to know, what kind of venues are in which district and, if he already was or lives in munich, wants to be able to move to a district, where the neighborhood is similar to the one he knows or were he lives in.

1.2 Solution

The KMeans clustering algorithm is used to cluster the neighborhoods of each district in munich according to its venues. This helps the user to identify similar districts and maybe to help him to find a district where he wants to live in.

2 Data

The postal code and district names of all districts of Munich are required to solve the task. The data published at `https://www.muenchen.de/int/en/living/postal-codes.html` is used in order to fetch the necessary data. An overview of the data from the page and therefore all districts and the according postal codes can be seen in table 1. As next step the latitude and longitude of all districts are required. They are then used for the foursquare API in order to get the venue data of all districts. So as first step the table 1 gets split, such that all postal codes have an own entry in the data table. Allach-Untermenzing has five different postal codes and therefore gets five entries in the new data table each with it's own postal code. This step is required, because each district can be very large and can therefore have a lot of different neighborhood clusters within it's district. The new table gets two additional columns. One containing the latitude and one the longitude. Table 2 contains the first five entries of this table. Afterwards this table gets used in order to get the venue data for each district by fetching the data using the foursquare API and the latitude and longitude values. In total 200 unique venue categories are available in Munich. Afterwards this data gets one hot encoded and grouped by each district and then a representative data set for each district is available for clustering.

	District	Postal Code
0	Allach-Untermenzing	80995, 80997, 80999, 81247, 81249
1	Altstadt-Lehel	80331, 80333, 80335, 80336, 80469, 80538, 80539
2	Au-Haidhausen	81541, 81543, 81667, 81669, 81671, 81675, 81677
3	Aubing-Lochhausen-Langwied	81243, 81245, 81249
4	Berg am Laim	81671, 81673, 81735, 81825
5	Bogenhausen	81675, 81677, 81679, 81925, 81927, 81929
6	Feldmoching-Hasenbergl	80933, 80935, 80995
7	Hadern	80689, 81375, 81377
8	Laim	80686, 80687, 80689
9	Ludwigsvorstadt-Isarvorstadt	80335, 80336, 80337, 80469
10	Maxvorstadt	80333, 80335, 80539, 80636, 80797, 80798, 8079...
11	Milbertshofen-Am Hart	80807, 80809, 80937, 80939
12	Moosach	80637, 80638, 80992, 80993, 80997
13	Neuhausen-Nymphenburg	80634, 80636, 80637, 80638, 80639
14	Obergiesing	81539, 81541, 81547, 81549
15	Pasing-Obermenzing	80687, 80689, 81241, 81243, 81245, 81247
16	Ramersdorf-Perlach	81539, 81549, 81669, 81671, 81735, 81737, 81739
17	Schwabing-Freimann	80538, 80801, 80802, 80803, 80804, 80805, 8080...
18	Schwabing-West	80796, 80797, 80798, 80799, 80801, 80803, 8080...

FIGURE 1 – Overview of Postal Codes and District Names in Munich

	District	Postal Code	Latitude	Longitude
0	Allach-Untermenzing	80995	48.190034	11.468105
1	Allach-Untermenzing	80997	48.192790	11.484461
2	Allach-Untermenzing	80999	48.195994	11.457013
3	Allach-Untermenzing	81247	48.176884	11.476058
4	Allach-Untermenzing	81249	48.176884	11.476058

FIGURE 2 – Table with latitude and longitude values for each postal code

3 Methodology

This section describes the algorithm further. As stated in section 2 the venue data for each district gets fetched from the foursquare API and is then used to cluster each neighborhood according to the available venues in a radius within 500 meters from the fetched latitude and longitude values. Then the data is explored further. The top five most common venues are printed for each district, after the table and it's venue categories were one hot encoded and grouped by the district names. This enables the kmeans algorithm to work with the data, because a Machine Learning algorithm is not able to work on textual data. The results can be seen in the according jupyter notebook. As next step the ten most common venues for each district are stored in a new data frame. This data frame helps to get an idea of the venues in each district. Figure 3 shows the first five rows of the resulting data frame. Afterwards the one hot encoded and grouped data is the input to the kmeans algorithm and the number of clusters is set to five. The resulting cluster labels are then additionally stored in the data frame containing the ten most common venues for each district.

	District	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	Allach-Untermenzing	Drugstore	Hotel	Trattoria/Osteria	Sporting Goods Shop	Yoga Studio	Falafel Restaurant	Fountain	Food Court	Food & Drink Shop	Food
1	Altstadt-Lehel	Supermarket	Sporting Goods Shop	Vietnamese Restaurant	Bus Stop	Light Rail Station	Bakery	Yoga Studio	Farmers Market	Fountain	Food Court
2	Au-Haidhausen	Bavarian Restaurant	Italian Restaurant	Food & Drink Shop	Beer Store	Tennis Court	Bakery	Donut Shop	Drugstore	French Restaurant	Fountain
3	Aubing-Lochhausen-Langwied	Gift Shop	Bus Stop	Trattoria/Osteria	Yoga Studio	Falafel Restaurant	Food Court	Food & Drink Shop	Food	Flower Shop	Fish Market
4	Berg am Laim	Playground	Pub	Rental Car Location	Light Rail Station	Yoga Studio	Event Space	Food & Drink Shop	Food	Flower Shop	Fish Market

FIGURE 3 – Overview of top ten venues for each district

4 Results and Discussion

This section shows the results of the kmeans clustering algortihm. In order to get a better understanding of the resulting clusters, a map of Munich and the clustered districts gets created. Figure ?? shows the final map from Munich. As one can see, the blue cluster is the most common cluster in Munich and therefore Munich seems to have a lot of similar districts in the city. The green cluster is more on the outer border of Munich and the other clusters are more distributed over the complete are of the city. So a user can now take a look on the districts and can compare the district he likes to others. This can help him finding a new apartment in a district he likes or to find activities in another neighborhood, which is similar to the one he likes and already knows.

5 Conclusion

The neighborhoods in Munich are now clustered an a map containing the results is available. Additionally, each cluster values are plotted in the according jupyter notebook in order to give the user more details of all clusters.