Capstone Project - The Battle of the Neighbourhoods (Week 2)

Applied Data Science Capstone by IBM
Completed by Md Wadud Hossain

Introduction

Munich is one of the populous cities of Germany. It is the capital of Bayern and one of the diverse cities in Germany. Munich is the home of Bayern Munich football club and also tourist hub of the beautiful alps dominant landscape. Moreover, yearly organizing city of famous Oktoberfest. It is also a global hub of business and commerce. Munich is also a city of two world famous university in the world and thousands of students from the different part of the world come here to study and research purposes.

Business Problem

As the number of residents increases every year, finding right place to live is always very difficult here, so is the finding of a good restaurants. Its very important to know that, whenever you are trying to move to new places, how is the new neighbourhood. What type of restaurants or supermarkets are around there?

Data and Methodology

I will be using https://www.muenchen.de/int/en/living/postal-codes.html for the postal code data and district name of Munich city to solve the task. To get the latitude and longitude data I will use python geopy library, where only name of the neighbourhood is required to find the latitude and longitude values for the given address.

[3]:

	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...
19	Schwanthalerhöhe	80335, 80339
20	Sendling	80336, 80337, 80469, 81369, 81371, 81373, 81379
21	Sendling-Westpark	80686, 81369, 81373, 81377, 81379
22	Thalkirchen-Obersendling-Fürstenried-Forstenri...	81379, 81475, 81476, 81477, 81479
23	Trudering-Riem	81735, 81825, 81827, 81829
24	Untergiesing-Harlaching	81543, 81545, 81547

Figure 1: Overview of all districts in Munich and their postal codes

After scraping the website, data will be stored in the data frames. In some districts, they have multiple postal codes! This is due to the largeness of all districts and to further divide them for getting a better localization. So as first step, each postal code gets it's own entry in a new pandas data frame in order to get more detailed information about the venues being in a small radius around the centre of each postal code.

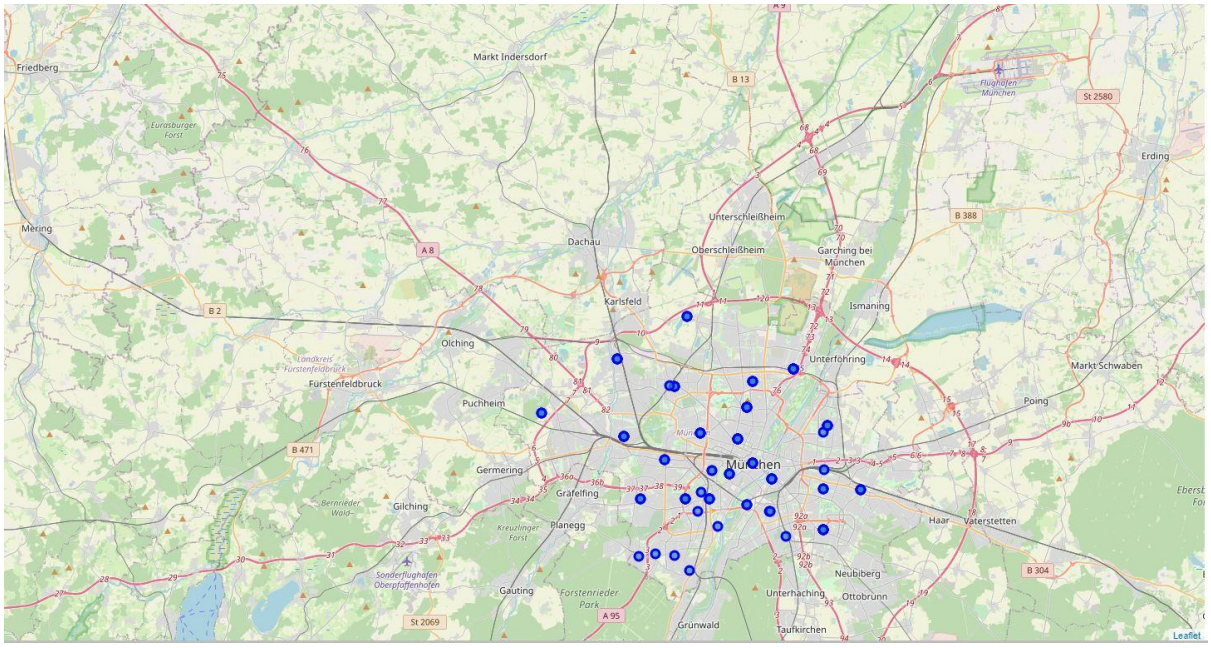

Figure 2: Map of districts in Munich split by their postal codes

As next step the available top 100 venues shall be fetched for each postal code. For this task, an API call to the Foursquare API is performed. The Foursquare API offers location data from all over the world for business purpose as well as for developers. I will also use Foursquare API. As Foursquare API offers location data from all over the world for business purpose as well as for developers.

The received venues for each district are stored within a new data frame having a shape of (3213, 7). Figure 4 shows the head of this new data frame.

In total there are 200 unique venue categories available. As next step the data has to be prepared for the clustering algorithm. In general, the K-Means algorithm only works with numerical data an we have categorical data. In order to be able to apply the K-Means algorithm, the Venue Categories first have to get one-hot encoded. Additionally, the one-hot encoded data frame gets grouped by the districts in order to have one row and therefore one cluster for each district.

```
[14]: munich_venues.head()
```

[14]:

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Allach-Untermenzing	48.195157	11.462973	Bäckerei Schuhmair	48.197175	11.459016	Bakery
1	Allach-Untermenzing	48.195157	11.462973	Sport Bittl	48.191447	11.466553	Sporting Goods Shop
2	Allach-Untermenzing	48.195157	11.462973	dm-drogerie markt	48.194118	11.465640	Drugstore
3	Allach-Untermenzing	48.195157	11.462973	Sicilia	48.193331	11.459387	Italian Restaurant
4	Allach-Untermenzing	48.195157	11.462973	Lidl	48.194428	11.465612	Supermarket

Figure 3: Head of new data frame containing the venues within each postal code district

Now the K-Means clustering algorithm gets applied. As input, the algorithm gets the data frame, containing the one-hot encoded venue categories, where the mean of the frequency of occurrence of categories was taken, and which was grouped by 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	Sporting Goods Shop	Italian Restaurant	Supermarket	Drugstore	Bakery	English Restaurant	Food & Drink Shop	Food	Fish Market	Fast Food Restaurant
1	Altstadt-Lehel	Drugstore	Sporting Goods Shop	Playground	Supermarket	Automotive Shop	Bakery	English Restaurant	Food	Fish Market	Fast Food Restaurant
2	Au-Haidhausen	Supermarket	Playground	Italian Restaurant	Automotive Shop	Drugstore	Bakery	Yoga Studio	English Restaurant	Food	Fish Market
3	Aubing-Lochhausen-Langwied	Drugstore	Italian Restaurant	Sporting Goods Shop	Electronics Store	Food	Fish Market	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Event Space
4	Berg am Laim	Supermarket	Drugstore	Automotive Shop	Yoga Studio	English Restaurant	Food & Drink Shop	Food	Fish Market	Fast Food Restaurant	Farmers Market

Figure 4: Top ten most common venues for each district

Finally, a map of all clusters gets created again by using the library Folium.

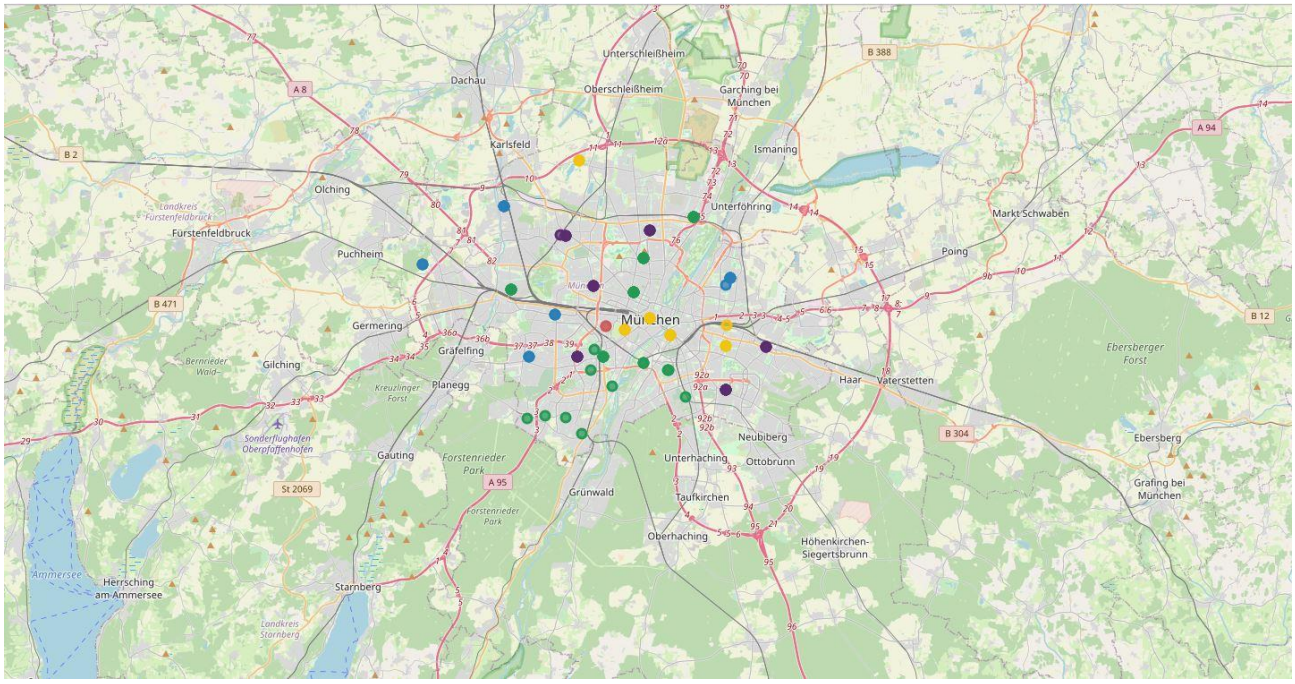

Figure 5: Map of clustered neighbourhoods in Munich

Results and Discussion

As we can see from the above data visualization, the green cluster in Munich are the most common clusters. Data suggest that Munich has lots of similar districts in it.

Though the green clusters are almost centre of the city compared to other clusters which are well distributed in the city. So, it will be very much easy for the users to learn more about the districts where they want to move next. Decision making would be more easier by using it. As we have also find out more about the clusters, we now know which cluster got the most common venues. So that you can understand the neighbourhoods more easily.

From the result, we can now call each cluster by their frequent venues and can predict the activities around. As result indicates, we can call the green cluster ‘Tourist cluster’, as it has hotels as most frequent venues and also plazas, we can easily estimate the activities in this neighbourhoods. You can probably find lots of advertisement in Airbnb for the short time stay.

Blue clusters are very good place to live, as you can find most frequently Restaurants and Bakery. Though I know most of the German love to live in the purple cluster where you can find German Restaurants, Irish Pubs and Café. We can easily call this cluster ‘German cluster’.

Though my favourite cluster is yellow cluster as you can see supermarkets are very frequent and lots of playground are around.

Conclusion

The purpose of this project was to compare the most popular venue categories and their respective venue types in Munich. So that a person from Munich will be able to make an educated decision as to which part of the city they would like to move next based on their own personal preferences.