Capstone Project - The Battle of Neighborhoods (Week 2)

„Giving recommendations for opening a new restaurant“

Report

1. Introduction / Business Problem

As part of my Capstone Project within this course I will make use of the secondly proposed idea in the module instructions. I will have a look on my home city Hamburg, Germany. I will use the Foursquare API for getting information about restaurants/cafe's in different neighborhoods. The goal is to give recommandations for opening a new restaurant and which kind of restaurant based on the density of them in the neighboorhood.

So who would be interested in this kind of project? Obviously it is someone who is looking to open a restaurant, but doesn't know if there is a lot of competition in the neighboorhood. If he/she is trying to start their own business it's a very relevant question. So the outcome of the project should lead in recommandations based on geographical and external data from Foursquare. Using some machine learning approaches data analytics will be executed and presented in geomaps using Folium.

2. Data Acquisition and cleaning

2.1 Data sources

As one data source I will of course use the Foursuqare API with all its relevant endpoints. Based on it, restaurant data will be gathered for all districts of the city. To get the district of the city I will scrape a table from Wikipedia. If there are more databases available, I will have a look on them, too.

2.2 Data cleaning

For example the table from Wikipedia, which contains information about the districts of the city, is not clean at all. There are different columns, which might not be relevant. On top of that many boroughs belong to the same district. So there is some work to do for cleaning the raw data.

2.3 Data preparation

Like shown in the previous modules of this course it is important to clean the data. But it's also required to prepare the data to adequate format, so it can be used easily in further steps. The Foursuare API is returnig a lot of information for example, from which some might not be interesting for the use case. That's why I should prepare the data, that they can support the goal of this project: show recommendations for opening a new restaurant based on geographical neighboorhod data.

3. Methodology and Exploratory Data Analysis

As mentioned before, I am using the Foursquare API for getting information about italian restaurants in the city of Hamburg. Therefore I am forming API queries, which are using the specific Fourquare Category ID for italian restaurants *(4bf58dd8d48988d110941735)*, the credentials and a calculated geolocations from the town hall in Hamburg. So the friend is looking for opening a new italian restaurant in the direct city of Hamburg.

At first we should have a look on what kind of data we will receive from the API. We are getting a JSON formatted response with all the nearby restaurants. Above you can see a sample output of the JSON response for one restaurant.

```
{
   'id':'4b5d8911f964a5207d6029e3',
   'name':'Ristorante Portonovo',
   'location':{
      'address':'Alsterufer 2',
      'lat':53.55922629808987,
      'lng':9.996573014390101,
      'labeledLatLngs': …
      'distance':993,
      'postalCode':'20354',
      'cc':'DE',
      'city':'Hamburg',
      'state':'Hamburg',
      'country':'Deutschland',
      'formattedAddress': …
   },
   'categories':[
      {
         'id':'4bf58dd8d48988d110941735',
         'name':'Italian Restaurant',
          …
      }
   ],
   'referralId':'v-1569899032',
   'hasPerk':False
}
```

For getting the results, that we are looking for in the end, I am using the passed geolocation of the restaurants for further preparation. By following the methodology the next step is to use the latitutde and longitude values for each restaurants to show their position on the map. You can see it on above picture. The restaurants are colored blue wheras the center (town hall) is displayed as a red circle. For that the package Folium was used.

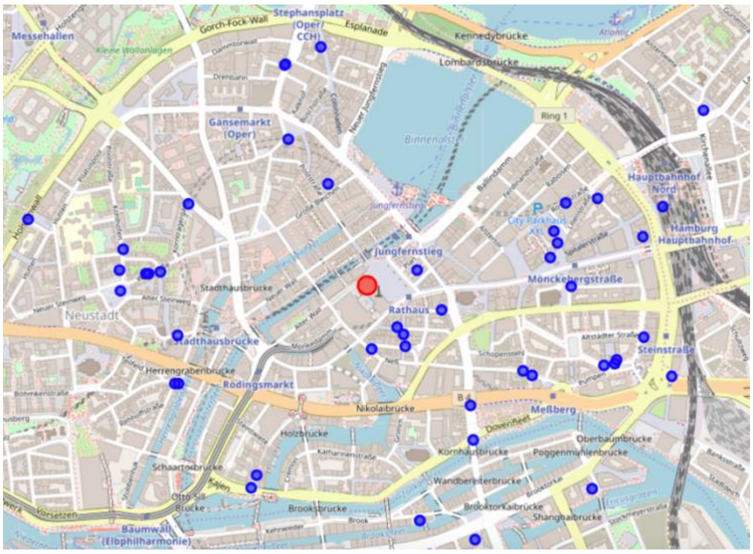

Figure 1: Map with italian restaurants in the city of Hamburg

For futher data exploration, besides basic visual stochastical approaches, I used K-Means clustering algorithm from Sklearn. As a data basis I am preparing the geolocation data for each restaurant in equivalent data arrays using numpy. We can use the latitude and longitude as our source data, because it can be used to identify a point on a chart and their axises. Therfore I am feeding the raw geodata of the restaurant into the K-Means Object. At this point I am using the method „.fit()“. For getting a great overview how the K-Means clustering algorithm performs, I tried different amount of clusters. In this case, I am showing two of the best: **5, 7**.

As an output of the Sklearn Class we will receive the centers of the identified clusters. In this case that are numerical values, which are usable for geographical purposes. So we can display those points on the Folium map as well.

In the above picture you can see the map from before with the centers of the clusters. They are indicated by a green colored circle. The *n_clusters* parameter in K-Means is set to 5.

In some other experiments I also changed the radius of the cluster center according to their area on the map.

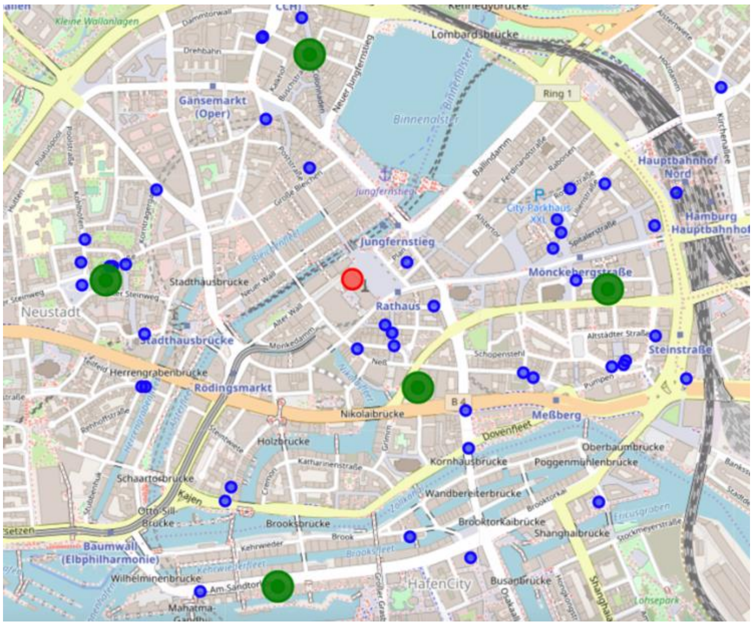

Figure 2: Map with restaurants and clusters centers (n=5, green)

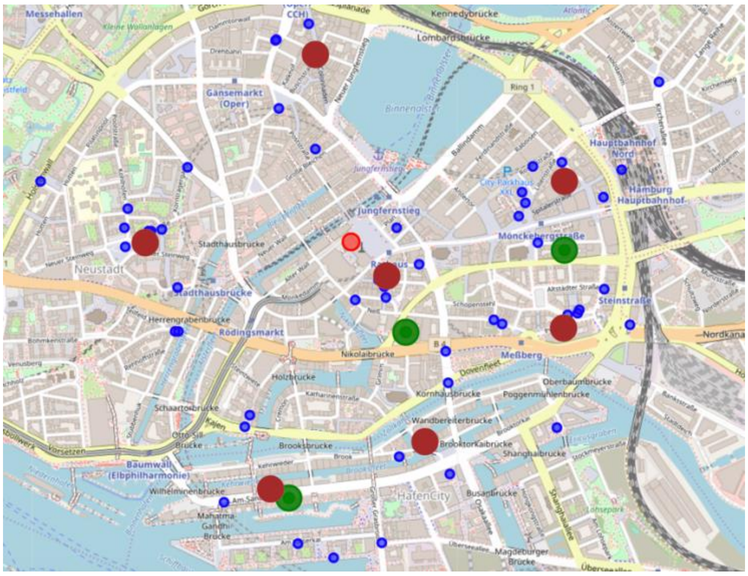

Figure 3: Map with restaurants and clusters centers (n=5, green | n=7, darkred)

4. Observations and Results

Based on the latest data exploration in previous chapter I can have some observations and results related to the goal to identify a good place for a new italian restaurant.

In figure 3 you can see all restaurants, all cluster centers for n=5 and n=7. We can see that the density of italian restaurants are more in the outter circles/areas of the city. The cluster centers are roughly representing the real density and location of italian restaurant groups. In the northern center there are almost no italian restaurants. There are many restaurants near the central train station (eastern direction). The cluster centers with an amount of 5 are even worse compared to the cluster centers with n=7. At least the most western and most northern cluster center haven’t changed regarding the amount.

5. Discussion and Conclusion

Now the big question is how we can use these obervations for giving a recommendation to our friend, who is looking for opening a new restaurant. You have to consider and to assume, that the prices for renting a place are really high in the direct center of the city. Therefore there are not that many italian restaurants. So it could be a hard job, to compensate the actual outlay / rental costs. Otherwise there are many people during the day, but most of the times occasional customers. So it coul be either a positive or negative point.

So by regarding the K-Means clusters on figure 3 it shouln’t be recommended to open a new restaurant at one of the direct cluster centers, because there are many italian restaurants already available.
It could be a recommendation to open a restaurant in the northern/western direction to the town hall. But also in between the to cluster centers at the left / middle position on the map.

Based on my simple analysis I can conclude that it could be good to open a new restaurant in between the fitted cluster centers and at their edges.