

Hotel Opening in Hamburg



Hamburg has everything, three rivers (Elbe, Alster and Bille) + two seas (North and Baltic). It has culture as a Free City and being the gateway to the New World. Worldly cosmopolitan and diverse city, Hamburg is one of Germany's finest tourist destinations.

Data:

1. We are using a .csv file containing all districts of Hamburg. We download it from the following source: <http://suche.transparenz.hamburg.de/dataset/stadtteil-profile>.
2. We are cleaning the data, constructing a data frame. Using the district names, we use the package `geopy.geocoders` to get the longitude and latitude values for each district. In order to keep the analysis short, we are just using the first 15 districts from our created data frame. In order to keep the analysis short, we are just using the first 15 districts from our created data frame.
3. Using the created data frame, we use the Foursquare API : (<https://developer.foursquare.com/docs/data>) in order to get data for the top venues in the different districts. Using the top venues we can analyze which district are the most popular and which districts have the most competition and ultimately we can say in which district we have the best chances for opening our new hotel.

Methodology:

For analyzing the topic, I used the following packages:

- Numpy (working with vectorized data)
- Pandas (for general data analysis)
- JSON (JSON library)
- Geocoders (Coordinates for the district of Hamburg)
- BeautifulSoup (Scraping html data, turns out I did not need at last)
- Matplotlib (for plotting)
- Sklearn (Cluster analysis)
- folium (for designing the map)

```

import numpy as np # vectorized data library

import pandas as pd # data analysis library
pd.set_option("display.max_columns", None)
pd.set_option("display.max_rows", None)
pd.options.mode.chained_assignment = None

import json # JSON library

from geopy.geocoders import Nominatim # convert an address into latitude and longitude values
!pip install geocoder
import geocoder # to get coordinates

import requests # library to handle requests
from bs4 import BeautifulSoup as bsoup # library to parse HTML and XML documents

from pandas.io.json import json_normalize # tranform JSON file into a pandas dataframe

# Matplotlib and associated plotting modules
import matplotlib.cm as cm
import matplotlib.colors as colors

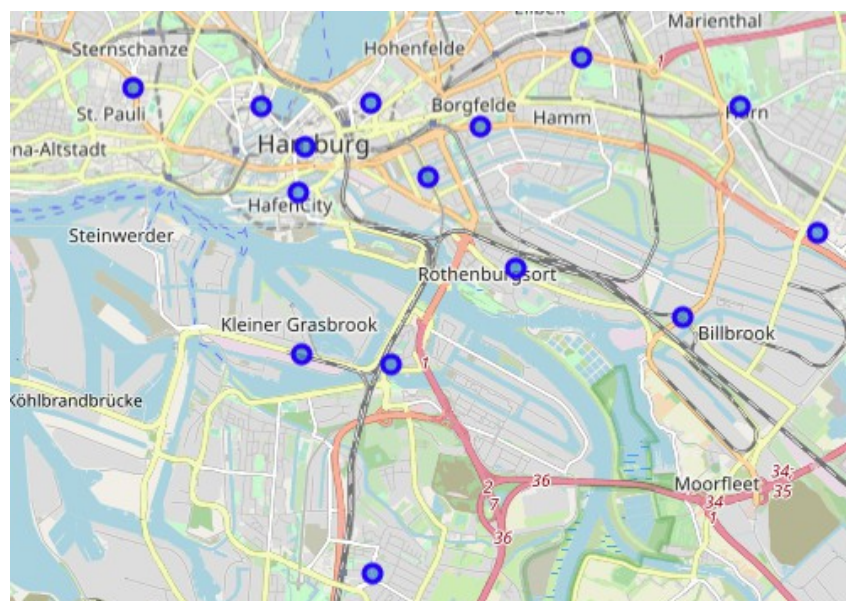
# import k-means from clustering stage
from sklearn.cluster import KMeans

import folium # map rendering library

print("Libraries imported.")

```

At first, BeautifulSoup should be used for scraping data. In the end a .csv file containing all districts of Hamburg was found. That turned out to be the simpler solution. The data was converted into a dataframe consisting just of the different districts of Hamburg. Using geocoder, the longitudinal and latitudinal data of these different districts were found and appended to the existing dataframe. Using folium a map of Hamburg and the different districts to be analyzed was created:



The next step was to access the Forsquare API. We need to get every venue of every different district in our dataframe. These results were grouped and appended to the existing dataframe. In order to be able to use ML algorithms onehot encoding was used. Each venue was categorized to a binary variable (0,1). To finish the analysis a new dataframe was created which only contained the district name, longitude,

latitude, and the top 10 venue of every district.

Because we want to find the perfect place for our hotel, kmeans algorithm seems to be the perfect method. By using kmeans algorithm we are able to compare the different districts and therefore be able to chose the perfect one for our new hotel.



In Picture No. 4 we can see the different created clusters. In the last analytical step, we list the results for every cluster. Now we are able to compare the different neighborhoods.

Result:

We wanted to open a hotel in a vivid district ('Viertel') of Hamburg.

Therefore we can now compare the formed clusters in order to find the perfect district for our business project.

1. Cluster 2 and 3 have the most entries of districts (4x). Therefore, we can assume that more people will be in those two clusters.

2. But we can see that for both clusters the top most venues are hotels (for Cluster 2 4 of 4, and for Cluster 3, 3 of 4). We can conclude that competition will be the highest in those two clusters. Therefore we should take a look at Cluster 1 and Cluster 4, which both have three districts. Cluster 4 has supermarkets as the top venues (3x) which is not the venues we are looking for. We are looking for restaurants, bar, cafes and entertainment venues in general. Cluster 1 does look better. We have a lot of restaurants, but also some hotels. That might be the best combination of chances and risks.

All in all, we would chose Cluster 1 based on the analysis. But a deeper and more detailed analysis is neccessary.

- Housing Prices?
 - Density of people and how many people are there in general?
 - Infrastructure?
- etc.