Applied Data Science Capstone Project

1. Introduction

The project consists of supporting the stakeholders of a small bakery franchise to extend their business in Hamburg. By leveraging the Foursquare location data, Hamburg neighborhood will be investigated based on their similarities, specifically in terms of availability of bakeries and cafes. In addition, the population distribution per Hamburg neighborhood, as well as, its population density will support the selection of the most promising region for the new business.

2. Data Description

Two sets of data were used in this project. The first set of data consists of information about Hamburg neighborhoods and their geographical location. Such information is available at https://www.geonames.org/postal-codes/DE/HH/hamburg.html.

These data will be scraped and combined with the population per Hamburg neighborhood data at https://de.wikipedia.org/wiki/Liste_der_Bezirke_und_Stadtteile_Hamburgs.

Finally, the districts of Hamburg will be classified based on their similarities using Foursquare location data and the most promising district will be selected based on their population information, as well as, the number of bakeries and cafes.

3. Data scraping and processing

The whole data extracting, scraping, processing, and analyses were performed using the Jupyter notebook and python 3.7.

The project was upload at Github and can be found at https://github.com/efeistauer/github-example/blob/master/Capstone_project_github.ipynb.

The list of libraries explored during this project is presented in Fig. 1.

```
#importing required libraries

import pandas as pd # library to process data as dataframes
from pandas.io.html import read_html

import numpy as np # library to handle data in a vectorized manner

import pandas as pd # library for data analsysis
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)

import json # library to handle JSON files

from geopy.geocoders import Nominatim # convert an address into latitude and longitude values

import requests # library to handle requests
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 matplotlib.pyplot as plt

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

#!conda install -c conda-forge folium=0.5.0 --yes # uncomment this line if you haven't completed the Foursquare API lab
import folium # map rendering library
```

Fig. 1: Libraries used during the capstone project

3.1 Scraping the data

After an extensive scraping of Hamburg's geographic data, a clean data frame with the information of each neighborhood postal code, latitude, and longitude was achieved. Fig. 2 present the head of the data frame. The tasks applied to clean the data are presented in the Github link with markdown describing each important step: https://github.com/efeistauer/github-example/blob/master/Capstone_project_github.ipynb.

	Neighborhood	Postal Code	Latitude	Longitude
0	Harburg	21073	53.459	9.981
1	Tonndorf	22045	53.587	10.117
2	Uhlenhorst	22085	53.571	10.019
3	Billstedt	22115	53.528	10.147
4	Langenhorn	22417	53.667	10.034

Fig. 2: Hamburg's neighborhood data

After scraping the data, a Hamburg map was created with Neighborhood superimposed on top to check the reliability of the data. Fig. 3 presents the superimposed data on the Hamburg map.

Fig 3. Neighborhood superimposed on top of Hamburg map

3.1 Collecting information for each Hamburg neighborhood

Based on the geographical location of each neighborhood, calls were made to the Foursquare API. Firs the code was validated for a neighborhood, were 100 venues within a radius of 1000 meters of Eppendorf were extracted.

After that, a function was created to repeat the same process and collect information about all Neighborhoods in Hamburg. The complete code is presented at jupyter notebook uploaded at https://github.com/efeistauer/github-example/blob/master/Capstone_project_github.ipynb.

Based on the information collected from the Foursquare API, the number of bakeries and cafes in Hamburg were filtered and stored a new data frame. Fig 4 presents the resultant data frame head.

	Neighborhood	Postal Code	Latitude	Longitude	Bakery	Café	Total
38	Hoheluft-Ost	20251	53.583	9.981	1.0	14.0	15
33	Altona-Altstadt	22767	53.55	9.935	6.0	6.0	12
9	Neustadt	20354	53.55	9.979	1.0	9.0	10
39	Hoheluft-West	20253	53.581	9.968	5.0	4.0	9
62	Sternschanze	20357	53.561	9.962	0.0	5.0	5

Fig 4. Number bakeries and cafes in Hamburg, extracted from Foursquare AIP.

3.2 Collecting data on the Hamburg population

The information about the Hamburg population and its distribution on each neighborhood was extracted from https://de.wikipedia.org/wiki/Liste_der_Bezirke_und_Stadtteile_Hamburgs.

After cleaning the data, a new data frame containing the Hamburg population per km2 was created (Fig. 5). Data scraping steps are described in the GitHub project: https://github.com/efeistauer/github-example/blob/master/Capstone_project_github.ipynb.

	Neighborhood	Area (km²)	Population	Population/km²
0	Hamburg-Altstadt	24	2350.000	979
1	HafenCity	22	4925.000	2239
2	Neustadt	23	12.762	5549
3	St. Pauli	25	22.097	8839
4	St. Georg	24	11.358	4733

Fig.5: Head of the data frame containing the population data per neighborhood

4. Results and discussion

The data bakeries and cafes information extracted from Foursquare API per Hamburg neighborhood were overlaid on the Hamburg map, as shown in Fig. 6. Each neighborhood was grouped according to the total quantity of establishments. Thus an interactive tool is provided for a easy visualization of establishments per neighborhood.

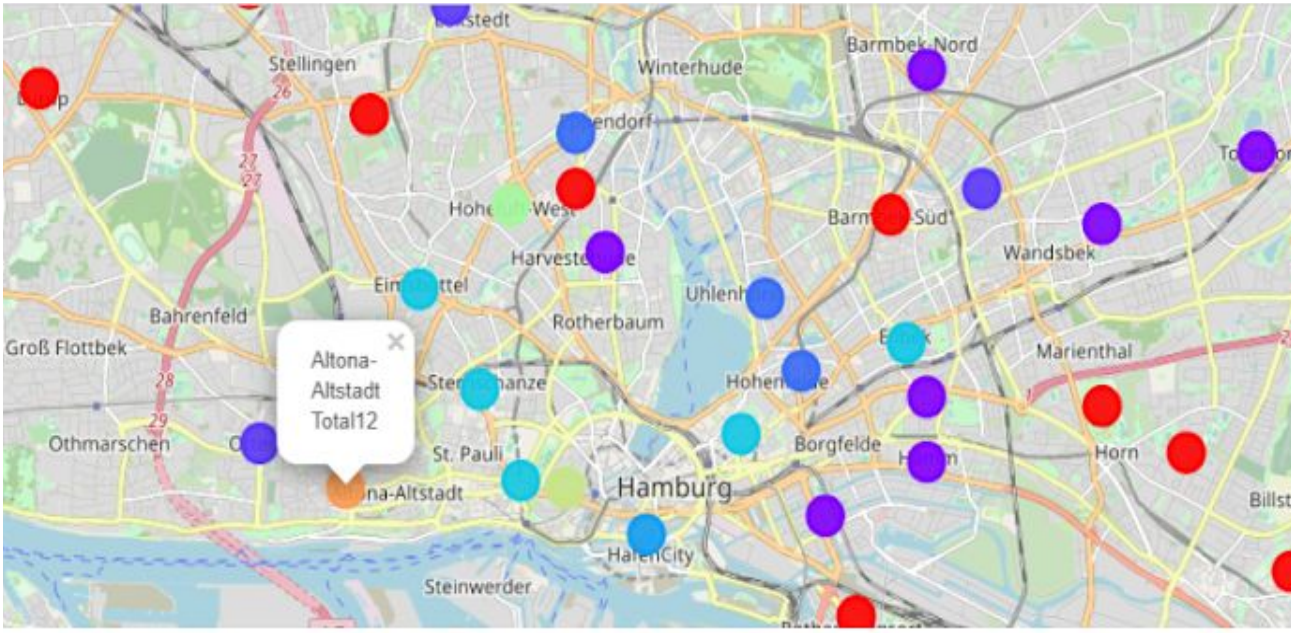

Fig 6. Number of establishment per Neighborhood superimposed on top of Hamburg map

Fig. 7 presents a graphic of 15 neighborhoods in Hamburg with more number of establishments, that is, the sum of bakeries and cafes. It was found out that Hoheluft-Ost, flowed by Altona-Altstadt and Neustadt presented the higher concentration of establishment, with cafes contributing more to the total number of establishments.

.

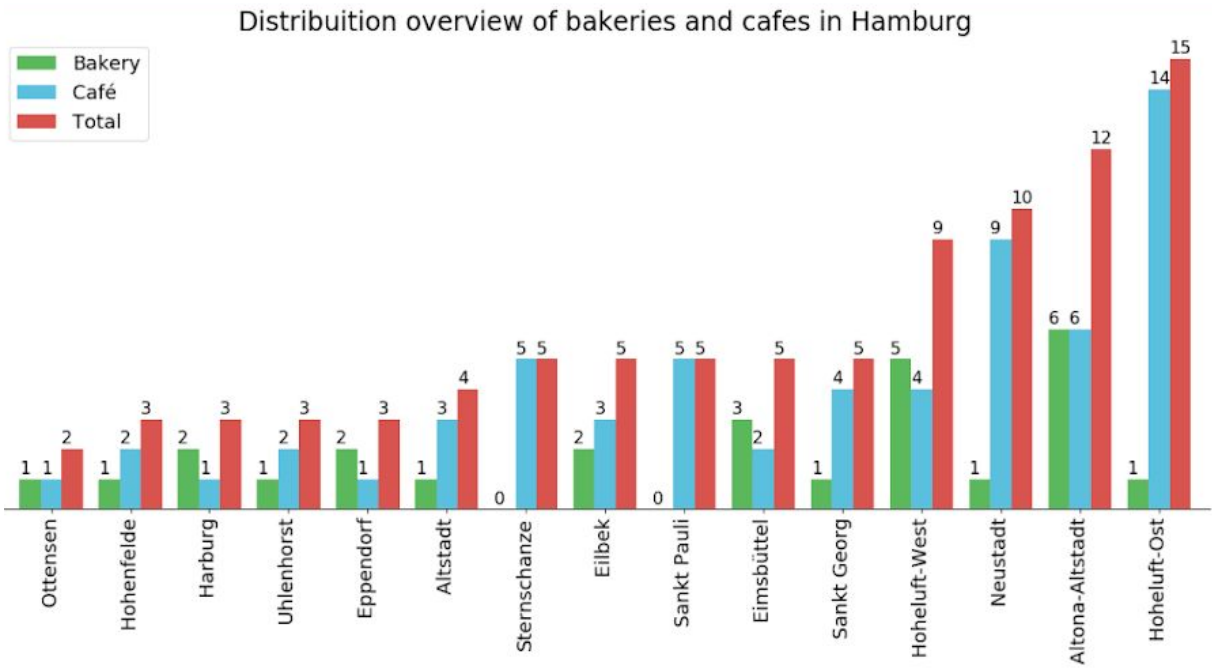

Fig. 7: 15 neighborhoods in Hamburg with more number of establishments

The number of establishments on its own does not provide enough information to decided on the most promising neighborhood to open a new business since it does not consider the population density per neighborhood. It is expected that the higher the population per km2 in a determined neighborhood the heigh is the amount of potential clients to the new business. Therefore, a graphic of the population density per neighborhood was created to support the assessment of the most promising region to open a new establishment, as shown in Fig.8.

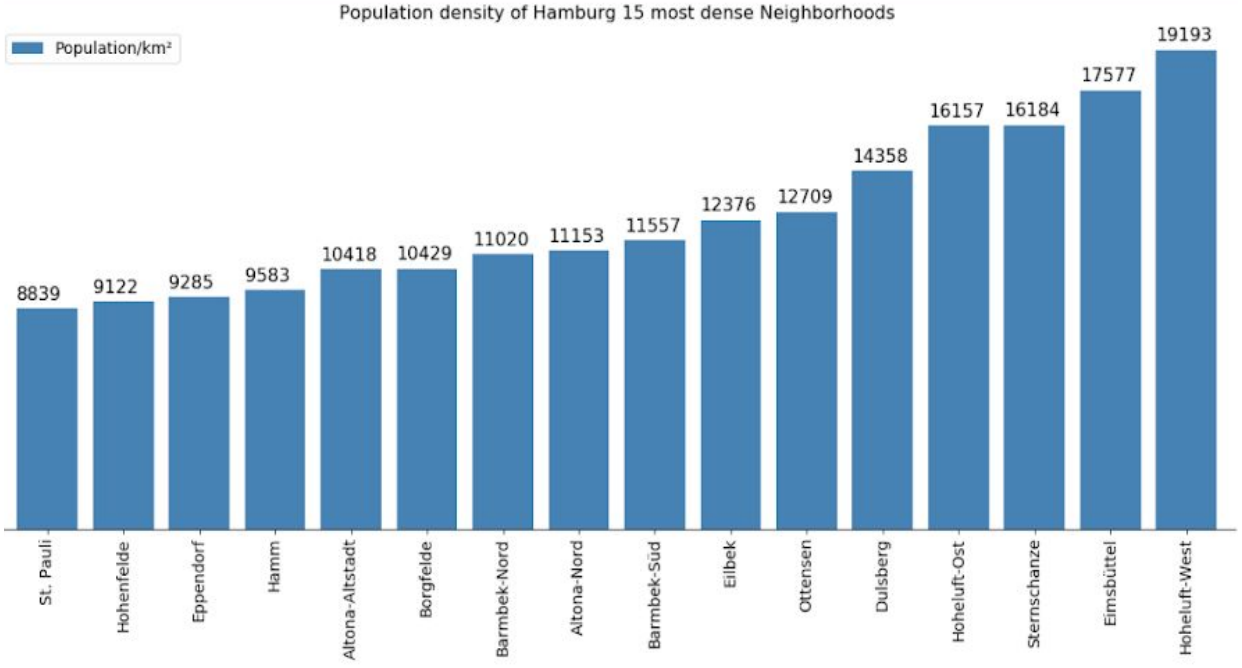

Fig. 8: 15 neighborhoods in Hamburg with higher density (population/km2)

Based on the two graphics one can easily observe that Neighborhoods such as Barmbek-Süd and Barmbek-Nor are between the densest neighborhood in Hamburg, however, they did not feature between the top 15 with most establishments in Hamburg (Fig. 7).

Such a simple assessment suggests that these neighborhoods might present a great opportunity for the extension of a bakery business. Besides that, Eppendorf one of the most preferred neighborhoods in the Hamburg feature between the 15 with higher density and presents only 3 establishments with only 1 cafe. Therefore, it might also accommodate more establishments.

5. Conclusions

This project successfully explored the Hamburg neighborhood and classified them based on the total quantity of available bakeries and cafes. The results were overlaid on the Hamburg map, providing an interactive graphic to visualize the distribution of establishment per neighborhood.

By comparing this data with the population density of each neighborhood, regions with a higher number of potential clients can be identified and support the decision on where to open a new establishment. Such an approach was applied first to bakeries and cafes, however, the available code provides information for 100 different venues per neighborhood.

The analysis showed that Barmbek-Süd and Barmbek-Nor are between the densest neighborhood in Hamburg, however, they did not feature between the top 15 with most establishments in Hamburg, suggesting that these neighborhoods might present a great opportunity for the extension of a bakery business. Besides that, Eppendorf one of the most preferred neighborhoods in the Hamburg feature between the 15 with higher density and presents only 3 establishments with only 1 cafe, which might indicate that it can accommodate more establishments.

It is important to mention, however, that such analysis did not consider the price involved on open a new establishment on each neighborhood, which is affected by different variables not investigated within this project.