EXPANSION TO GERMANY

Coursera Capstone Final Project

1. INTRODUCTION/BUSINESS PROBLEM

The client, a small event-organiser company in Copenhagen, Denmark has had a great last couple of years and is now looking to expand its business to Germany. Being a small company, it must spend its limited resources carefully even after the good years, they can only choose one city to start off with.

They want to know which city they should choose and ,potentially, some indication of where in the chosen city they should open a new office. Based on available resources, the company's analysts have narrowed down its expansion scope of interest to the two closest German major-cities of Hamburg and Berlin.

Having had some experience they know that they tend to do well and know how to market in neighbourhoods of Copenhagen, and they suspect it has something to do with the overall atmosphere in them. The "atmosphere" can be proxied by the types of venues that are most common in the neighbourhood.

The task then is to come up with a model prototype of determining the type of neighbourhood in the three cities , and give a better look of where the company is likely to do better in order for them to make the final expansion decision.

1.1 PROBLEM FORMULATION

So, the **business problem** can be formulated in brief as:

"Which of the two cities(Berlin, Hamburg) should the Danish event-organiser company expand to? "

And, the **data science problem** that is derived from the above can be formulated in brief as:

"Between Hamburg and Berlin, which city's neighborhoods are most similar to those of Copenhagen in terms of entertainment venues?"

1.2 COMMERCIAL INTEREST

Naturally, finding places that are very similar to the one that the client has already succeeded in, will increase chance of success in the new location. This type of analysis would allow the company to open to new markets utilizing resources they currently have more effectively, instead of going through large and potentially costly changes within the company's resources/knowledge base or having to rely purely on luck by expanding blindly.

2. DATA

2.1. DATA CHOICE AND SOURCES

The problem requires breaking the cities down into respective parts, then using the Foursquare API, to collect data om the most popular venues in each neighborhood for the analysis to determine their 'entertainment atmosphere'.

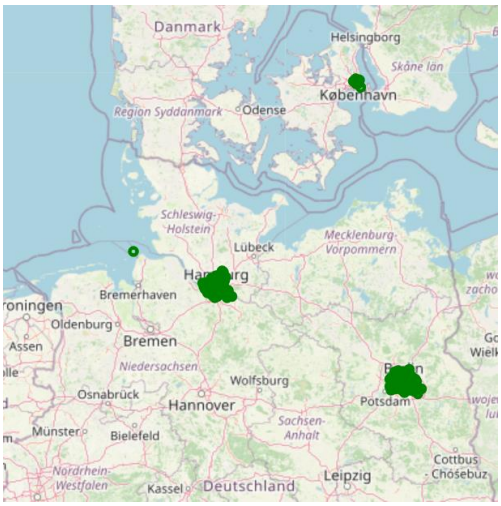
For Copenhagen, there are 10 neighborhoods scraped from the internet, 9 neighborhoods(with 'Amager' being considered as one) corresponding to Copenhagen and one for Frederiksberg, a legally independent municipality that is geographically still part of the City of Copenhagen. While these neighborhoods can be broken down further into areas, this will not be done in the prototype, because only this way can the neighborhoods of Copenhagen be comparable to neighborhoods of the two other cities(due to the latter being significantly larger in size and population).

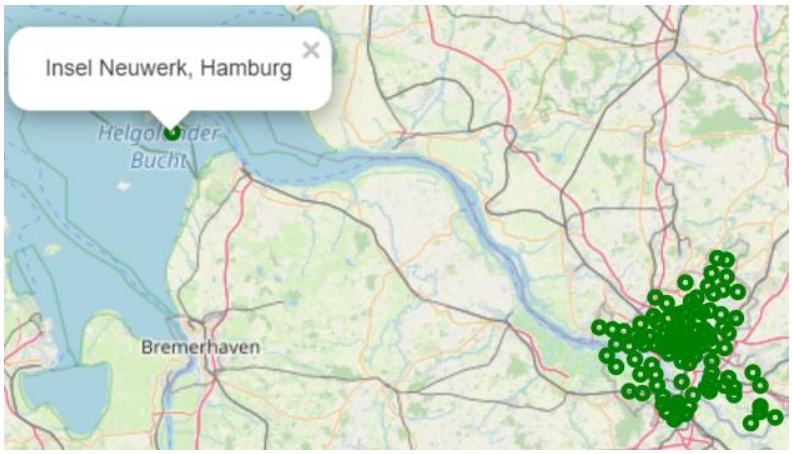
Hamburg is broken down into 7 boroughs which can then be further broken down into 104 quarters, which can be scraped from the links provided. The quarters are the neighborhoods which can then form a basis for comparison to the ones of Copenhagen. It must be noted that the neighborhood of 'Insel Neuwerk'(see figures 1,2) will be excluded as it is an island quarter that belongs to Hamburg but is an outlier neighborhood because it is ~120km away from the Hamburg city center, and it would be impractical for the client to travel to or build an office in. There are also administrative neighborhoods for people who live on boats, but this will also be excluded for outlier reasoning, and keep the notion of 'neighborhoods' in the traditional sense.

Berlin, the capital of Germany, is composed of 12 boroughs, that can be subdivided into 96 neighborhoods, the data is going to be scraped from the given website.

The data for geospacial coordinates is going to be obtained via. GeoPy Nominatim API.

It must be noted, that population will not be used in the models of this prototype, but only to get the reader more familiar with the cities in question via EDA and descriptive statistics. Population and size may play a part in expansion decisions but what value the supposed client would place on these factors is subjective, the analysis is focused on finding the better option for expansion via similarity using venue data. The interested reader can extend this model further applying more data as they see appropriate, including their own 'size-similarity' trade-off.





FIGURES 1,2: LOCATIONS OF INTEREST; 'INSEL NEUWERK'

Table summary of the data and sources

DATA	SOURCE
COPENHAGEN	https://www.citypopulation.de/en/denmark/copenhagen/
HAMBURG	https://www.citypopulation.de/en/germany/hamburg/admin/
BERLIN	https://en.wikipedia.org/wiki/Boroughs_and_neighborhoods_of_Berlin
VENUES	https://api.foursquare.com/v2/venues/
GEO SPECIAL DATA(API)	https://nominatim.org/