EXPANSION TO GERMANY

Coursera-Capstone Final Project Report

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 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 neighbourhoods 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 the 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 neighbourhood for the analysis to determine their 'entertainment atmosphere'.

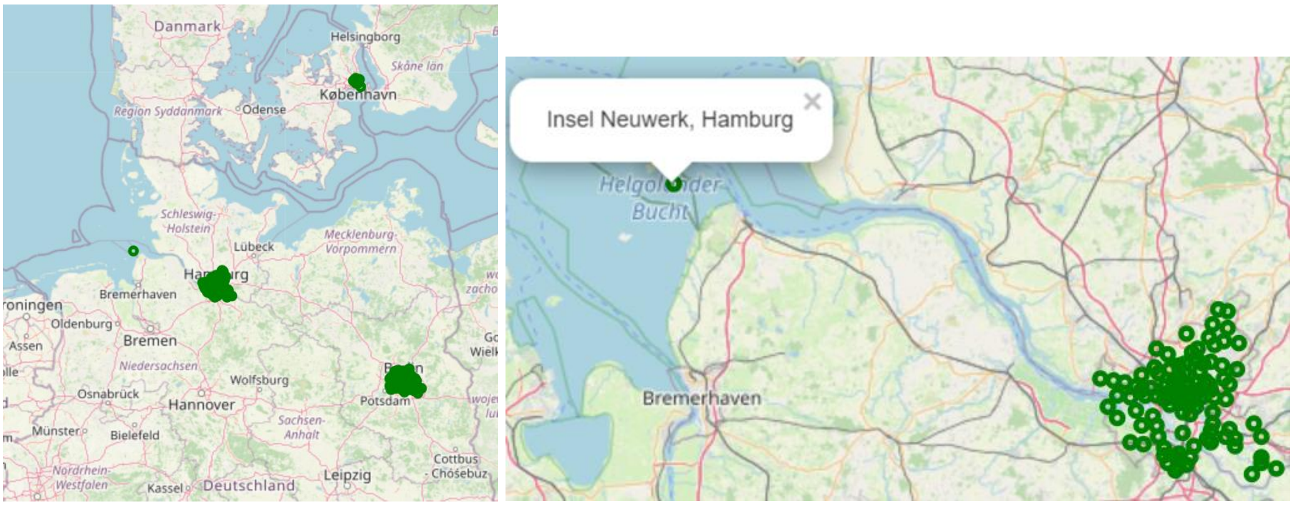
For Copenhagen, there are 10 neighbourhoods scraped from the internet, 9 neighbourhoods (with 'Amager' being considered as one) corresponding to Copenhagen and one for Frederiksberg, a legally independent municipality that is geographically still part of the City of Copenhagen. While these neighbourhoods can be broken down further into areas, this will not be done in the prototype, because only this way can the neighbourhoods of Copenhagen be comparable to neighbourhoods 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 neighbourhoods which can then form a basis for comparison to the ones of Copenhagen. It must be noted that the neighbourhood of 'Insel Neuwerk'(see figures 1,2) will be excluded as it is an island quarter that belongs to Hamburg but is an outlier neighbourhood because it is ~120km away from the Hamburg city centre, and it would be impractical for the client to travel to or build an office in. There are also administrative neighbourhoods for people who live on boats, but this will also be excluded for outlier reasoning, and keep the notion of 'neighbourhoods' in the traditional sense.

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

The data for geospatial coordinates are 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 by applying more data as they see appropriate, including their own 'size-similarity' trade-off.



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

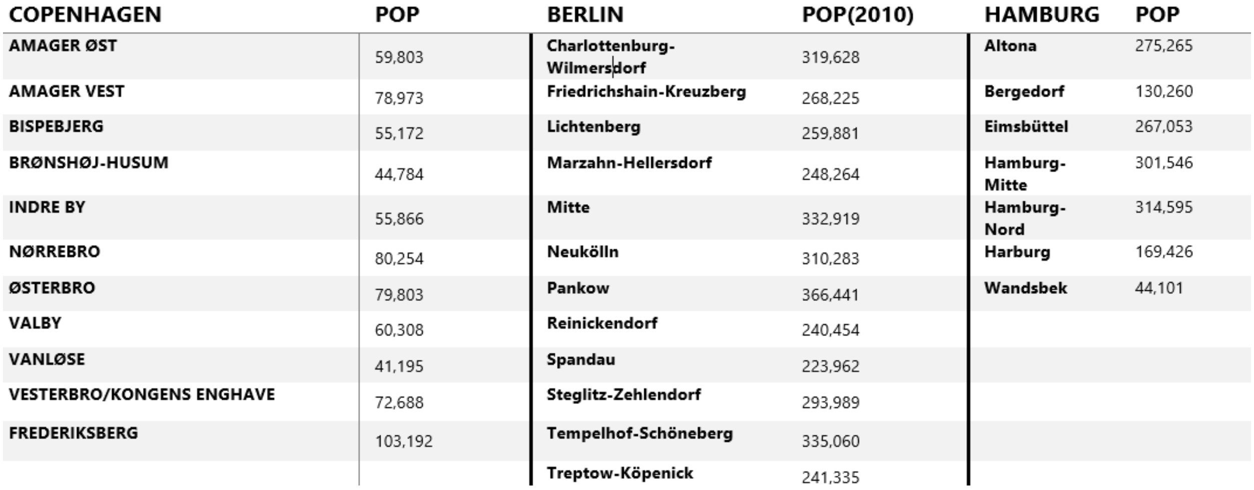
Table summary of the data and sources:

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/

3. METHODOLOGY

3.1 EDA AND DESCRIPTIVE STATISTICS OF THE NEIGHBOURHOOD AND BOROUGHS POPULATIONS

Below are the populations of the boroughs of the three cities for the reader to get a better sense of the cities used in the analysis. Unsurprisingly, Copenhagen considering that Germany is far more populous than Denmark. While the borough data is manageable, both Hamburg and Berlin have around 100 neighbourhoods, and the interested reader can check the notebook,



COPENHAGEN	POP	BERLIN	POP(2010)	HAMBURG	POP
AMAGER ØST	59,803	Charlottenburg-Wilmersdorf	319,628	Altona	275,265
AMAGER VEST	78,973	Friedrichshain-Kreuzberg	268,225	Bergedorf	130,260
BISPEBJERG	55,172	Lichtenberg	259,881	Eimsbüttel	267,053
BRØNSHØJ-HUSUM	44,784	Marzahn-Hellersdorf	248,264	Hamburg-Mitte	301,546
INDRE BY	55,866	Mitte	332,919	Hamburg-Nord	314,595
NØRREBRO	80,254	Neukölln	310,283	Harburg	169,426
ØSTERBRO	79,803	Pankow	366,441	Wandsbek	44,101
VALBY	60,308	Reinickendorf	240,454		
VANLØSE	41,195	Spandau	223,962		
VESTERBRO/KONGENS ENGHAVE	72,688	Steglitz-Zehlendorf	293,989		
FREDERIKSBERG	103,192	Tempelhof-Schöneberg	335,060		
		Treptow-Köpenick	241,335		

TABLE 1: CITY BOROUGH POPULATIONS

The boxplot below summarises the borough population data of the three cities, while Berlins boroughs in Berlin are most populous, Hamburg has the Borough with the largest population of all three cities, while boroughs of Copenhagen have considerably fewer people.

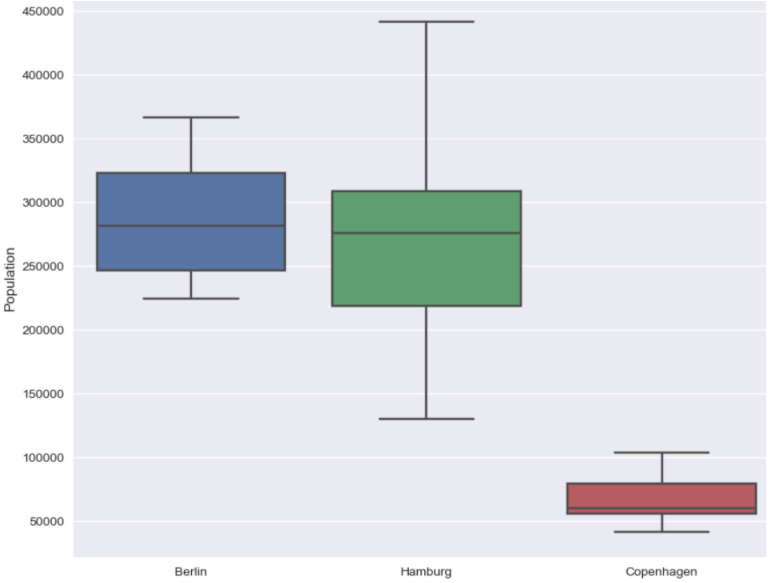


FIGURE 3: BOXPLOT BOROUGH POPULATIONS

Figure 4 shows the distributions of the populations in the neighbourhoods of the three cities. Berlin, understandably, has most populous neighbourhoods on average, while Hamburg has a lot of rather small neighbourhoods of less than 10,000 people. The histogram overlay also shows that treating the Copenhagen Municipalities as neighbourhoods are justified since they are right in the middle of the population values overall and breaking them down further might have damaged location comparability.

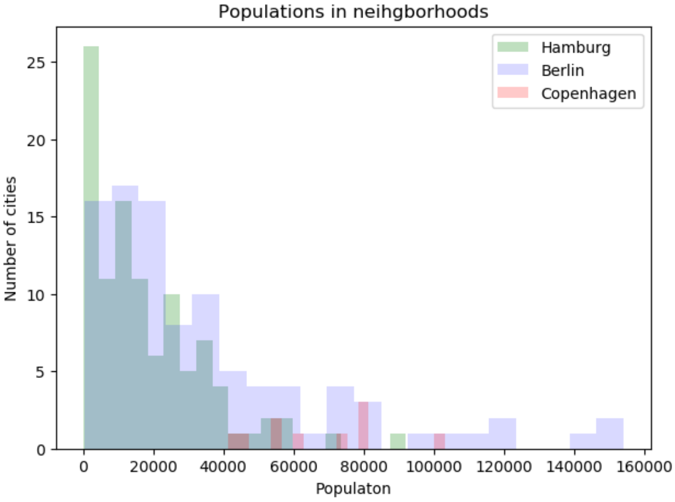
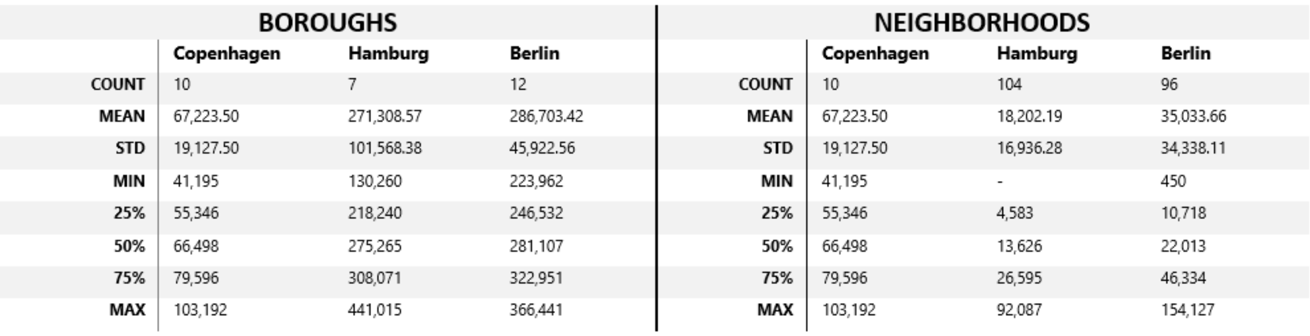


FIGURE 4: NEIGHBORHOOD HISTOGRAM OVERLAY

A numeric summary of the visualisations is provided in Table 2.

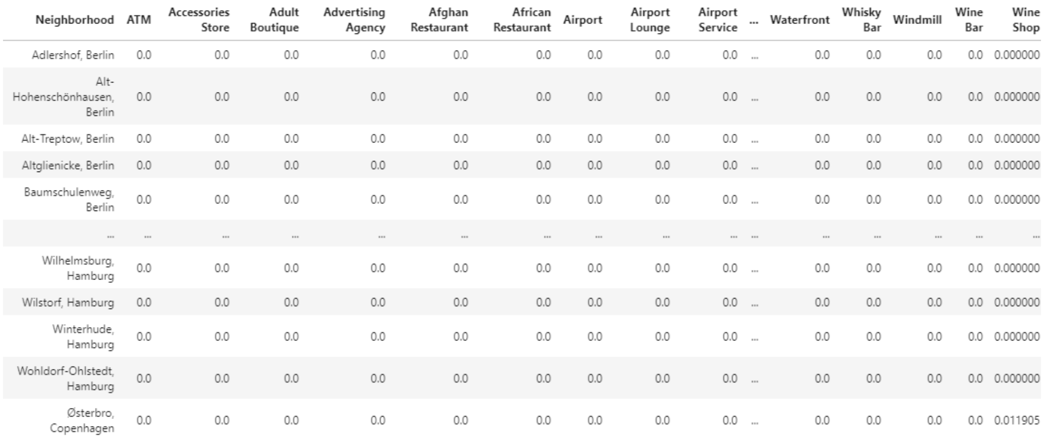


	BOROUGHS				NEIGHBORHOODS		
	Copenhagen	Hamburg	Berlin		Copenhagen	Hamburg	Berlin
COUNT	10	7	12	COUNT	10	104	96
MEAN	67,223.50	271,308.57	286,703.42	MEAN	67,223.50	18,202.19	35,033.66
STD	19,127.50	101,568.38	45,922.56	STD	19,127.50	16,936.28	34,338.11
MIN	41,195	130,260	223,962	MIN	41,195	-	450
25%	55,346	218,240	246,532	25%	55,346	4,583	10,718
50%	66,498	275,265	281,107	50%	66,498	13,626	22,013
75%	79,596	308,071	322,951	75%	79,596	26,595	46,334
MAX	103,192	441,015	366,441	MAX	103,192	92,087	154,127

TABLE 2: SUMMARY DESCRIPTIVE STATISTICS OF POPULATIONS

A snippet of the basis of the "entertainment atmosphere" of the neighbourhoods' Foursquare data is shown in table 3. Unfortunately, even in the prototype there are 8516 different venues in 416 categories (expressed with 'One-hot encoding' in the algorithm) neither can be broken down into effective descriptive visualisations or tables so the interested reader is referred to the notebook to see the details.

The dataset of 216 neighbourhoods is exemplified in Table 3, the numbers represent the proportion of the venue type of all gathered venues, for example, there are 84 total gathered venues for Østerbro, Copenhagen(bottom right corner) and one of them is 1 'Wine Shop' so wine shops represent ~1.12%(= 1/84) of the gathered venues in the area. This 'proportionality' of relative representation of the venue type will be the basis for the neighbourhood 'entertainment atmosphere' similarity estimation(see subsection 3.2).

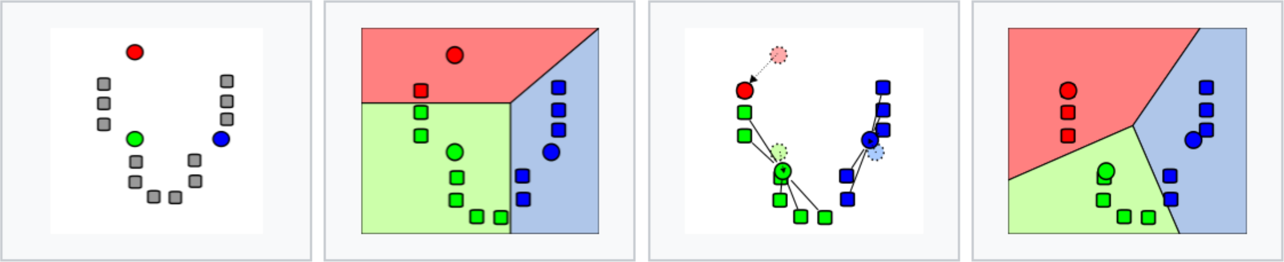


Neighborhood	ATM	Accessories Store	Adult Boutique	Advertising Agency	Afghan Restaurant	African Restaurant	Airport	Airport Lounge	Airport Service	...	Waterfront	Whisky Bar	Windmill	Wine Bar	Wine Shop
Adlershof, Berlin	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.000000
Alt-Hohenschönhausen, Berlin	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.000000
Alt-Treptow, Berlin	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.000000
Altglienicke, Berlin	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.000000
Baumschulenweg, Berlin	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Wilhelmsburg, Hamburg	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.000000
Wilstorf, Hamburg	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.000000
Winterhude, Hamburg	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.000000
Wohldorf-Ohlstedt, Hamburg	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.000000
Østerbro, Copenhagen	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.011905

TABLE 3: SNIPPET OF VENUE DATA

3.2. Estimation method

A “K-means clustering algorithm” will be used to calculate the similarity of the neighbourhoods. K-means algorithm labels N-dimensional points, where N stands for feature dimension of each point, by, putting K number of centroids between and assigning a label to each x_i by based on the closest centroid. The algorithm then tries to optimize for square distance by changing the position of the centroids until no point changes label. A basic illustration is given in Figure 5.



1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3. The centroid of each of the *k* clusters becomes the new mean.

4. Steps 2 and 3 are repeated until convergence has been reached.

FIGURE 5: SCHEMA FOR K-MEANS(SOURCE, HTTPS://EN.WIKIPEDIA.ORG/WIKI/K-MEANS_CLUSTERING)

In the prototype, the Euclidian distance will be used. Five centroids will be picked(K=5) and will be initially picked at random. It is important to note that results will vary dependent on the initiation since a local optimum will be found, with no guarantees for it being global, for this reason, and reproducibility it should be noted that results are obtained through random seed 77.

In our case, the points x_i are the neighbourhoods of the three cities (see Table 3), N corresponds relative representation 416 one-hot encoded venues, in exactly the way of Table 3, described in the previous subsection. The algorithm will then use the venue representation to cluster the neighbourhoods and label the 0-4.

The result will show to which city the client should expand to based on which German city has the most label overlap (i.e. label similarity) with Copenhagen. For example, if Copenhagen has 7 out of 10 neighbourhoods belonging to cluster 1, Hamburg has 80 in cluster 1, and Berlin has 13, Hamburg is then preferable to our client since it is more similar and will be easier to expand to.

4. RESULTS

The algorithm for the dataset has converged in 11 iterations with a total sum of square distances (inertia) of ~12.285. The overall distribution of neighbourhoods is given in Figure 6. Cluster label 3 has the most neighbourhoods overall with 112, while cluster 2 seems to represent a neighbourhood that is an outlier potentially but could also mean that the number of clusters is too high in the prototype.

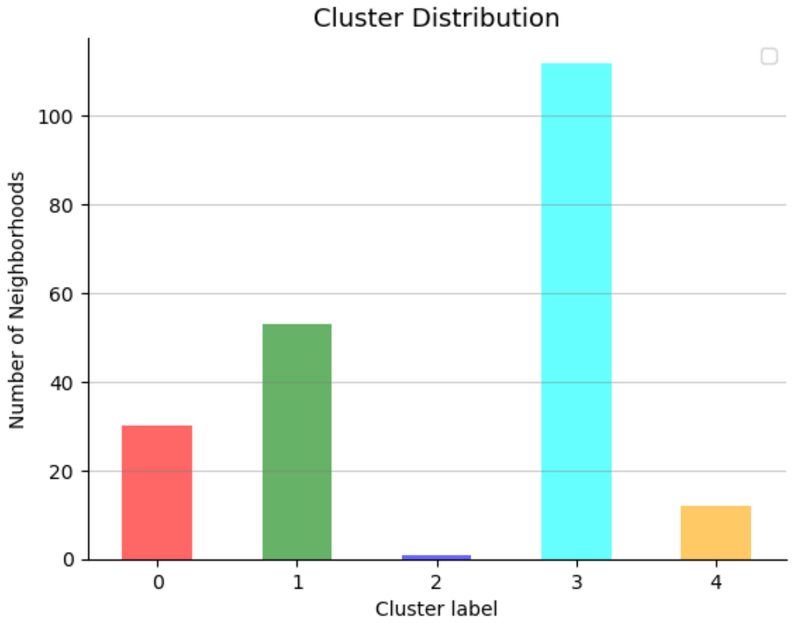


FIGURE 6: CLUSTER LABEL DISTIBUTION

CLUSTER LABEL	NUMBER OF NEIGHBORHOODS
3	112
1	53
0	30
4	12
2	1

TABLE 4: CLUSTER DISTRIBUTION

Figure 7 shows the classification of neighbourhoods in each city. The client's home of operations, Copenhagen is completely homogenous with all the neighbourhoods being clustered under label 3. Berlin and Hamburg have a relatively close number of cluster 3 neighbourhoods with 50 and 52 respectively. Hamburg has most of cluster 0, 4 and 2(outlier) type neighbourhoods, while Berlin has the majority of cluster 1.

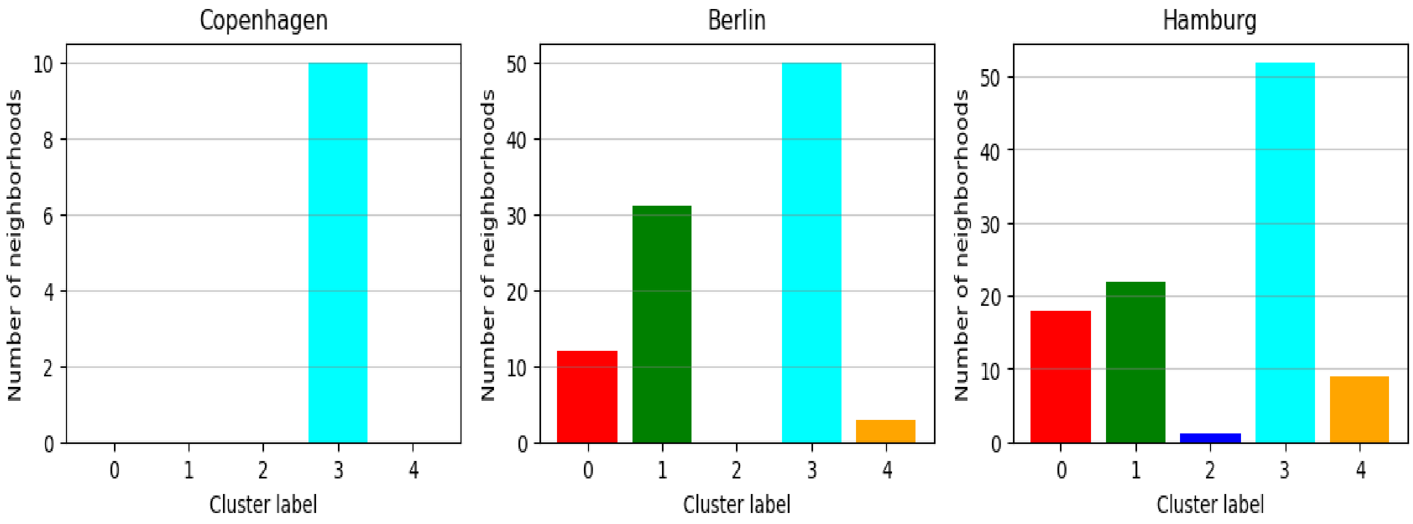
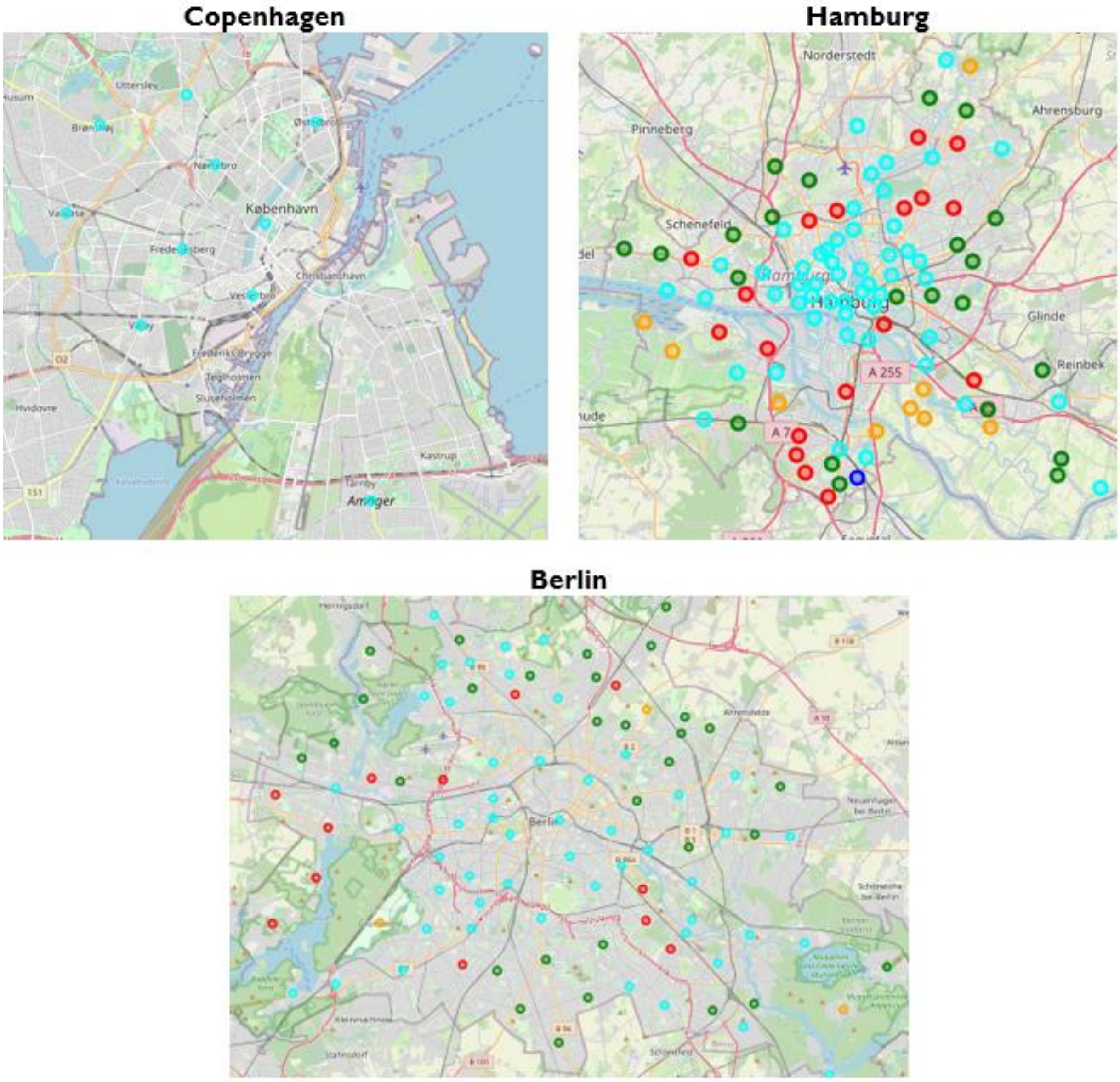


FIGURE 7: CLUSTER DISTRIBUTIONS BY CITY

To infer the possible nature of the clusters maps of the three cities is given in figures 8-10. We can see that for the German cities Cluster 3(Cyan) represents the city centre locations, so the chosen neighbourhoods of Copenhagen are the most similar to central Berlin and Hamburg. Cluster 1(Green)objects seem to appear more often in the edges of the city and could represent industrial exit areas since they tend to appear next to major motorways. Cluster 0(Red) look like suburban family areas while the patterns for cluster 4 seem to be hard to discern.



Copenhagen

Hamburg

Berlin

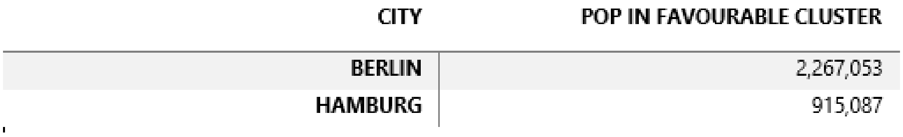
FIGURES 8-10: CLUSTERS MAPPED

5. Discussion

The above analysis showed that the client is best suited at targeting central locations of both German cities. From figures 8-10 we can see that were to choose Berlin the prototype algorithm would advise them to stick to the Western part of Berlin ideally centrally such as the areas of, Charlottenburg, Westend or Halensee.

Were the client to chose Hamburg they are better off sticking to Northern parts of Hamburg according to the algorithm ideally around the northern branch of river Elbe or south of the Outer Alster, so areas like and close to Borgfelde, Hohenfelde or Harvested.

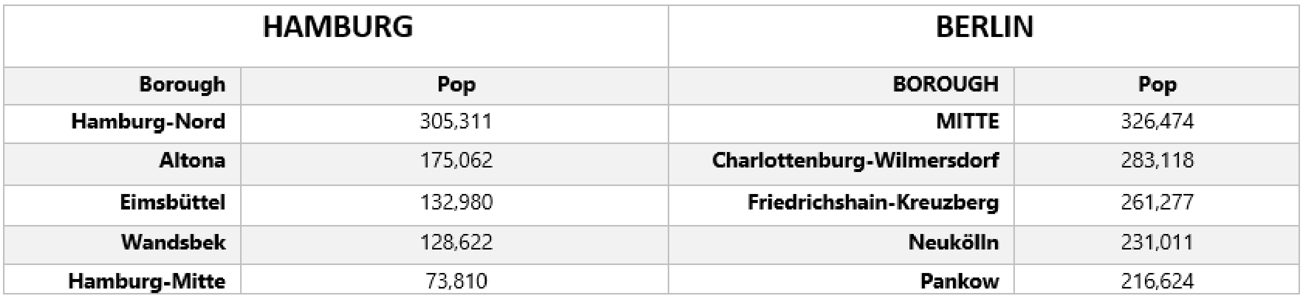
The choice of the city could depend on the population in the clusters. The table shows the overlapping Cluster 3 in Berlin and Hamburg, Berlin has overall more people in neighbourhoods that are like those in Copenhagen, however, the population of neighbourhoods in Hamburg(915,085) is closer to that of Copenhagen(~700,000).



CITY	POP IN FAVOURABLE CLUSTER
BERLIN	2,267,053
HAMBURG	915,087

TABLE 5: POPULATION IN CLUSTER 3 NEIGHBORHOODS IN THE GERMAN CITIES

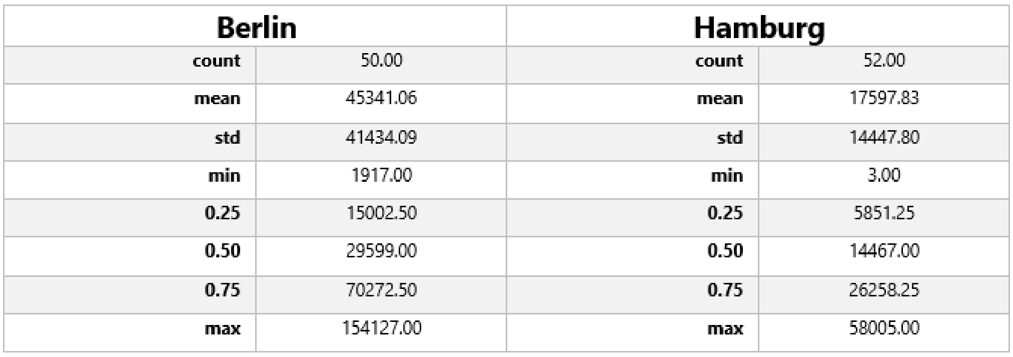
To see the best boroughs to open the office by population Top 5 boroughs based on favourable(Cluster 3) population are given in Table 6. Interestingly, Northern Hamburg is the best borough to open the office(as inferred from the map inspection before), and not central as in Berlin, this fact may potentially weigh into the decision, offices in city centres are usually more expensive, but this is outside of the scope of this report.



HAMBURG		BERLIN	
Borough	Pop	BOROUGH	Pop
Hamburg-Nord	305,311	MITTE	326,474
Altona	175,062	Charlottenburg-Wilmersdorf	283,118
Eimsbüttel	132,980	Friedrichshain-Kreuzberg	261,277
Wandsbek	128,622	Neukölln	231,011
Hamburg-Mitte	73,810	Pankow	216,624

TABLE 6: TOP5 BOROUGHS BY FAVOURABLE NEIGHBORHOOD POPULATION

Finally, the favourable neighbourhood population distribution descriptions are shown in Table 7. While there were a few more favourable neighbourhoods in Hamburg the average and median population of them was considerably smaller.



Berlin		Hamburg	
count	50.00	count	52.00
mean	45341.06	mean	17597.83
std	41434.09	std	14447.80
min	1917.00	min	3.00
0.25	15002.50	0.25	5851.25
0.50	29599.00	0.50	14467.00
0.75	70272.50	0.75	26258.25
max	154127.00	max	58005.00

TABLE 7: DESCRIPTIVE STATISTICS OF POPULATIONS OF FAVOURABLE NEIGHBORHOOD

5.1 Potential limitations and their resolution of the prototype

There are a few issues and how to deal with them must be discussed in the prototype model for completeness and honesty.

Firstly, the choice of hyper-parameters. The number of clusters and the type of distance was rather arbitrary, for the number of clusters one can set up a reasonable range of models and use metrics like inertia(see section 4) can then be used to determine the optimal number of clusters.

Secondly, there is no guarantee of a global optimum, and there is inherent randomness to the model depending on where the initial centroids are placed. This issue could be potentially be solved or at least improved by running a model a relatively large number of times(say, 2000) and picking the model with the smallest inertia, but even then there are no guarantees of a global optimum.

Finally, there is the issue of model accuracy. The prototype uses only one type of data to determine clusters, that is, only the venue data from Foursquare API. This limitation can be overcome with more or better data sources. Other relevant data (e.g. socio-economic data) can be analogously applied, rather easily to this prototype algorithm, one just needs to research and justify any features being added.

6. CONCLUSION

Overall, there are more people in neighbourhoods of Berlin that are similar to those in Copenhagen, while Hamburg has overall more similar neighbourhoods. The best locations(Boroughs) to open the office are Central(Mitte) Berlin, and Northern Berlin.

Nevertheless, there are plenty of people in either city and the client themselves needs to decide whether to pick Berlin with the potentially harsher competition or Hamburg with potentially less business. Given that the client does pick either of the two cities the prototype(maybe with some of the improvements from section 5.1) will aid them to a more successful and efficient expansion.