Applied Data Analysis Capstone Project
Buying a home in Copenhagen

Gesine Wanke

July 12, 2020

1 Introduction

This report will present the analysis of housing prices and their increase over the last five years in the city districts of Copenhagen. It will further analyse the venues of the neighborhoods around the public transport system of Copenhagen's Metro as it allows to easily combine biking with other means of commuting, which is most common in Copenhagen. This project will allow people to evaluate an offer for a future home in terms of investment as well as in terms of venues that are important for the individual customer.

1.1 Background

Buying a place to live is an important decision in peoples lives. Firstly, financially buying a new home is a huge commitment. The home should therefore be affordable and not be overpriced compared to the housing prices in the area. From a financial perspective an area where house prices can expected to increase in future would be preferred.
Secondly, many other aspects of living have to be considered when finding a new home, such as how far it is to commute to work or schools, or how well the neighborhood is connected via public transport to the rest of the city. As a place to live people have different preferences for venues surrounding their home. As the attractiveness of each area is perceived differently by individual people, a similarity analysis between different neighborhoods will make it faster to find alternative neighborhoods that can be considered for future homes.
This project will make it easier to decide for the area of buying a home in Copenhagen as well as comparing an offer for a home to the average pricing and the financial perspective of the area.

1.2 Target audience

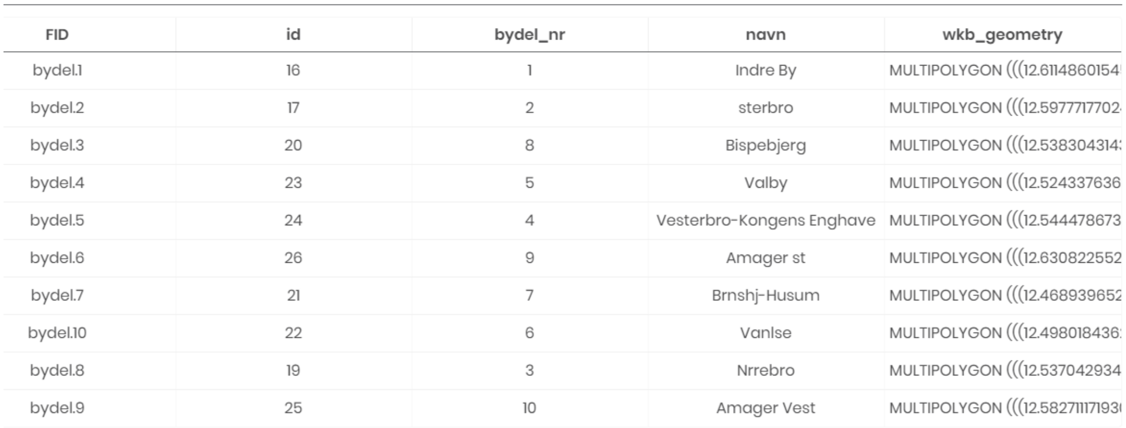
The analysis is targeted for people that are in the process of buying or considering to buy an apartment in Copenhagen. It aims to allow an easier and faster decision of where to search for a new home financially, but also from the aspect of attractiveness of a neighborhood. It further allows the customer to evaluate an explicit offer in regards of the average prices of apartments in the area, as well as the expected value increase in future for the place in question.

2 Data

The following section describes the used data and gives examples for the used data sets. The Data for the analysis is collected from several sources. The borders of the city districts are accessible form the danish open-source data base opendata.dk (`https://www.opendata.dk/city-of-copenhagen/bydele`). Price histories of housing prices are available for the Copenhagen city districts from the danish organisation Boliga (`https://www.boliga.dk/boligpriser`). For the data of the public transport system a list of metro stations is extracted from Wikipedia (`https://en.wikipedia.org/wiki/List_of_Copenhagen_Metro_stations`). A link to the individual stations is used to access Wikipedia's data of longitude and latitude of the metro stations. The Foursquare data base (`https://foursquare.com`) is used to extract venues and their respective category in the neighborhood of the metro stations.

2.1 Copenhagen district borders

The .geojson files for the districts of Copenhagen are retrieved from opendata.dk (`https://www.opendata.dk/city-of-copenhagen/bydele`) and are downloaded within the python script. The file contains the borders as latitude and longitude for each district. These are used to define the borders of the coropeth map that will show the housing prices and the price increases. Figure 1 shows the dataset as it can be downloaded from the webpage.



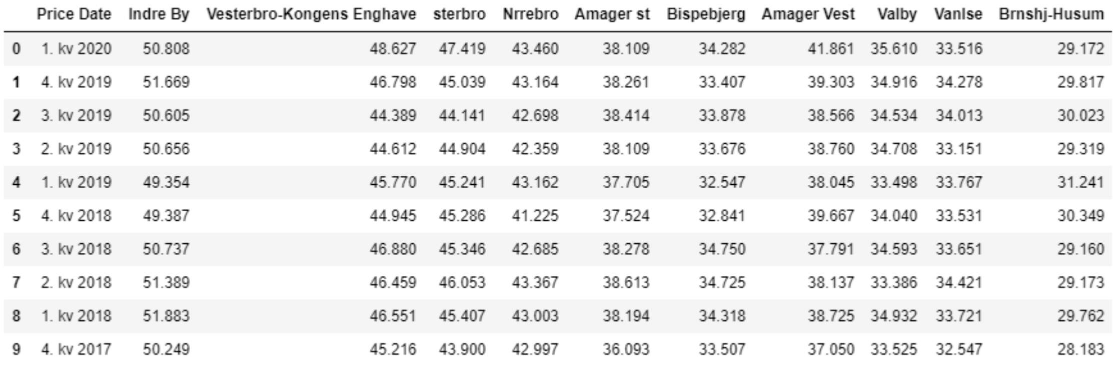
FID	id	bydel_nr	navn	wkb_geometry
bydel.1	16	1	Indre By	MULTIPOLYGON (((12.6114860154
bydel.2	17	2	sterbro	MULTIPOLYGON (((12.5977717702
bydel.3	20	8	Bispebjerg	MULTIPOLYGON (((12.538304314
bydel.4	23	5	Valby	MULTIPOLYGON (((12.524337636
bydel.5	24	4	Vesterbro-Kongens Enghave	MULTIPOLYGON (((12.544478673
bydel.6	26	9	Amager st	MULTIPOLYGON (((12.630822552
bydel.7	21	7	Brnshj-Husum	MULTIPOLYGON (((12.468939652
bydel.10	22	6	Vanlse	MULTIPOLYGON (((12.498018436
bydel.8	19	3	Nrrebro	MULTIPOLYGON (((12.537042934
bydel.9	25	10	Amager Vest	MULTIPOLYGON (((12.5827111719

Figure 1: City district data in a geojson-file from opendata.dk (`https://www.opendata.dk/city-of-copenhagen/bydele`)

The key "navn" will be used to link the city districts to the data of the analysis for plotting the coropeth maps. The "wkb_geometry" column contains the longitude and latitude of the borders of the city districts.

2.2 Copenhagen housing prices

The housing prices and housing price history are available for each city district of Copenhagen from Boliga's hompeage (`https://www.boliga.dk/boligpriser`). The extracted data is the price per square meter of an appartment over the last 5 years. The data can not scraped directly but the relevant data is copied to a .csv-file to be read into a pandas data frame from the git-hub repository. The file contains the name of the district, as well as the price history of the per square meter. Figure 2 shows the extracted housing prices per square meter for the districts of Copenhagen in 1000 DKK. The "Price Date" column shows the date of the price data, under the neighborhood columns the price per square meter in 1000 DKK for apartments can be found.

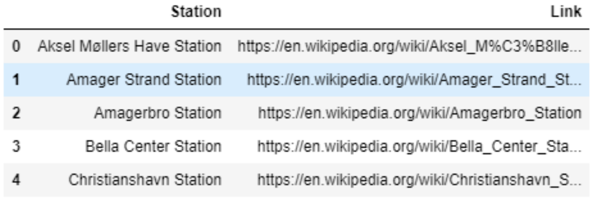


	Price Date	Indre By	Vesterbro-Kongens Enghave	sterbro	Nrrebro	Amager st	Bispebjerg	Amager Vest	Valby	Vanlse	Brnshj-Husum
0	1. kv 2020	50.808	48.627	47.419	43.460	38.109	34.282	41.861	35.610	33.516	29.172
1	4. kv 2019	51.669	46.798	45.039	43.164	38.261	33.407	39.303	34.916	34.278	29.817
2	3. kv 2019	50.605	44.389	44.141	42.698	38.414	33.878	38.566	34.534	34.013	30.023
3	2. kv 2019	50.656	44.612	44.904	42.359	38.109	33.676	38.760	34.708	33.151	29.319
4	1. kv 2019	49.354	45.770	45.241	43.162	37.705	32.547	38.045	33.498	33.767	31.241
5	4. kv 2018	49.387	44.945	45.286	41.225	37.524	32.841	39.667	34.040	33.531	30.349
6	3. kv 2018	50.737	46.880	45.346	42.685	38.278	34.750	37.791	34.593	33.651	29.160
7	2. kv 2018	51.389	46.459	46.053	43.367	38.613	34.725	38.137	33.386	34.421	29.173
8	1. kv 2018	51.883	46.551	45.407	43.003	38.194	34.318	38.725	34.932	33.721	29.762
9	4. kv 2017	50.249	45.216	43.900	42.997	36.093	33.507	37.050	33.525	32.547	28.183

Figure 2: Head of data frame showing price history for apartment prices per sqare metre in 1000 DKK from `https://www.boliga.dk/boligpriser`.

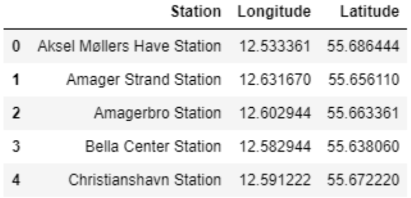
2.3 Copenhagen metro stations

For the analysis of the neighborhoods in the districts and their venues the metro-system of Copenhagen is used. The stations provide easy access to public transport for commuting within the city, especially when combined with biking. The metro stations are scraped from a wikipedia-list (`https://en.wikipedia.org/wiki/List_of_Copenhagen_Metro_stations`). From the list the station name, as well as the link to the wikipedia entry of each respective station is extracted. Figure 3 shows an example of the extracted data from the list of metero stations. The link to each Wikipedia entry is used to extract the location data for the metro stations from the pages. Three metro stations have been neglected as no location data could be extracted from the link or no Wikipedia page existed. Figure 4 shows an example of the extracted geo-data for the metro stations.



	Station	Link
0	Aksel Møllers Have Station	https://en.wikipedia.org/wiki/Aksel_M%C3%B8lle...
1	Amager Strand Station	https://en.wikipedia.org/wiki/Amager_Strand_St...
2	Amagerbro Station	https://en.wikipedia.org/wiki/Amagerbro_Station
3	Bella Center Station	https://en.wikipedia.org/wiki/Bella_Center_Sta...
4	Christianshavn Station	https://en.wikipedia.org/wiki/Christianshavn_S...

Figure 3: Head of data frame showing the metro stations and links extracted from Wikipedia `https://en.wikipedia.org/wiki/List_of_Copenhagen_Metro_stations`.

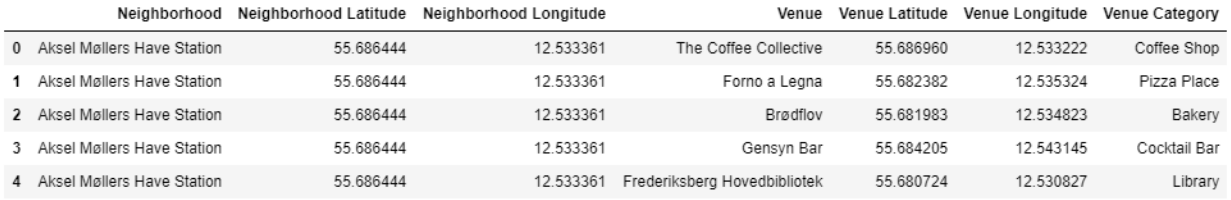


	Station	Longitude	Latitude
0	Aksel Møllers Have Station	12.533361	55.686444
1	Amager Strand Station	12.631670	55.656110
2	Amagerbro Station	12.602944	55.663361
3	Bella Center Station	12.582944	55.638060
4	Christianshavn Station	12.591222	55.672220

Figure 4: Head of data frame showing the longitude and latitude of the Copenhagen metro stations.

2.4 Copenhagen venues

The Fourthsqare data base (`https://foursquare.com`) is used to extract a list of venues around 2km within a radius of each metro station. This distance is chosen on the one hand as it is easily biked in Copenhagen, which is a very common thing to do. On the other hand, it guarantees that a sufficiently large number of venues can be found so the neighborhoods around the stations can be compared by a similarity analysis. Figure 5 shows an example of the extracted data from Foursquare for Axel Møllers Have Station. The data frame contains the neighborhood data for the station (name, latitude and longitude) as well as the venues, their respective geo-data and the category of the venue.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Aksel Møllers Have Station	55.686444	12.533361	The Coffee Collective	55.686960	12.533222	Coffee Shop
1	Aksel Møllers Have Station	55.686444	12.533361	Forno a Legna	55.682382	12.535324	Pizza Place
2	Aksel Møllers Have Station	55.686444	12.533361	Brødflov	55.681983	12.534823	Bakery
3	Aksel Møllers Have Station	55.686444	12.533361	Gensyn Bar	55.684205	12.543145	Cocktail Bar
4	Aksel Møllers Have Station	55.686444	12.533361	Frederiksberg Hovedbibliotek	55.680724	12.530827	Library

Figure 5: Head of data frame showing the list of venues for the metro station Axel Møllers Have Station extracted from `https://foursquare.com`.

3 Methodology

The following describes the methods used for the analysis of the data. For the project the housing prices are analysed and the neighborhoods around the metro stations are compared with similarity analysis.

3.1 Explanatory analysis

The data analysis is performed within a jupyter notebook and python 3.6 running on IBM Watson Studio cloud. The main libraries for the data analysis are numpy, pandas and sklearn. Folium is the major library used for the visualisation of the results on a map of Copenhagen.

3.1.1 Housing prices

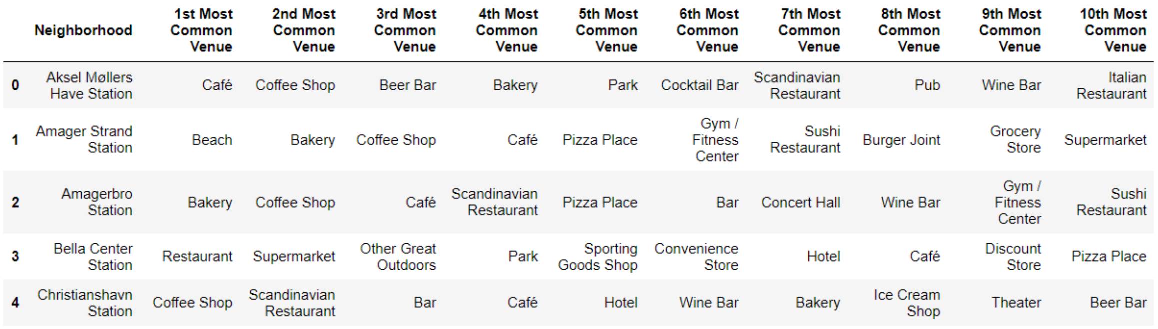
For the housing prices the average annual change is calculated. For average square metre price for the year 2019 $\bar{P}_{2019}$ and 2015 $\bar{P}_{2015}$ are calculated for each city district. The average annual price increase ΔP from the last five years is calculated as

$$\Delta P = \frac{\bar{P}_{2019} - \bar{P}_{2015}}{\bar{P}_{2015}} \cdot \frac{100\%}{5} \tag{1}$$

Both, the current housing square metre price (available for the first quarter 2020) and the average annual price increase are visualised as coropeth plots with the folium library. Thus, the geo-data for the city district borders are used to define the districts on the map. The pricing data is linked via the name of the district to the map.

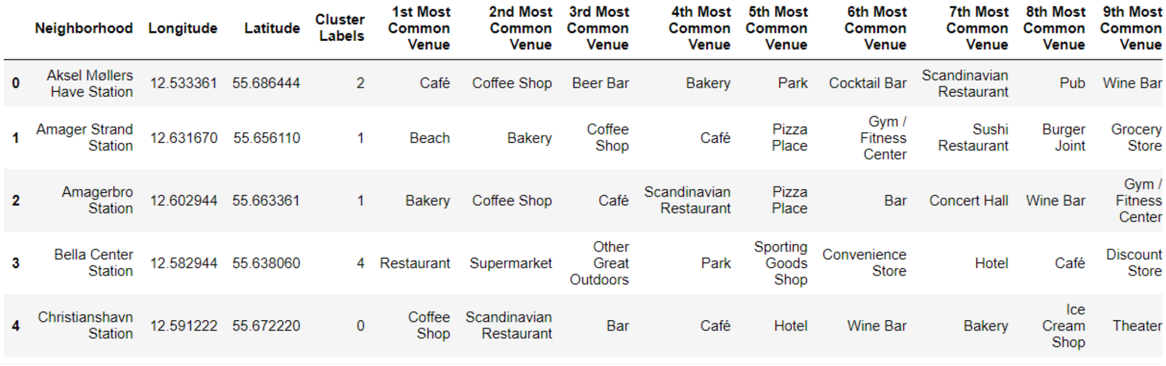
3.1.2 Clustering of neighborhoods

With the Foursquare API the list of venues around each Metro station in Copenhagen is extracted as a representation of the different neighborhoods. The radius around each station is 2km, and the number of venues extracted is limited to 100 per station. A relatively large radius is chosen as it incorporates an area that is commonly biked in Copenhagen. It also assures that a sufficient number of venues is extracted for each station allowing for a similarity analysis. It has been checked that more than 70 venues are available in each neighborhood. For each neighborhood a frequency count of the category of the venues is done. From the frequency count the 10 most common venues can be joined in a data frame. Figure 6 shows the head of the resulting data frame as an example. The machine learning algorithm ”K-means” from the sklearn library is used to cluster neighborhoods that are similar to each other. In total 5 groups of clusters are used. The cluster number is joined with the geo-data and the frequency count of the venues for each station in a data frame. Figure 7 shows the head of the resulting data frame, here for visualisation purposes truncated to nine most common venues. The metro stations are plotted on top of the coropeth map of house prices via their geo data. The markers are colored by the cluster number the station is assigned to bey the k-means algorithm. The markers contain pop-up labels showing the 5 most common venues of the neighborhood. This visualization allows an easy consideration of alternative neighborhoods for buying a home.



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Aksel Møllers Have Station	Café	Coffee Shop	Beer Bar	Bakery	Park	Cocktail Bar	Scandinavian Restaurant	Pub	Wine Bar	Italian Restaurant
1	Amager Strand Station	Beach	Bakery	Coffee Shop	Café	Pizza Place	Gym / Fitness Center	Sushi Restaurant	Burger Joint	Grocery Store	Supermarket
2	Amagerbro Station	Bakery	Coffee Shop	Café	Scandinavian Restaurant	Pizza Place	Bar	Concert Hall	Wine Bar	Gym / Fitness Center	Sushi Restaurant
3	Bella Center Station	Restaurant	Supermarket	Other Great Outdoors	Park	Sporting Goods Shop	Convenience Store	Hotel	Café	Discount Store	Pizza Place
4	Christianshavn Station	Coffee Shop	Scandinavian Restaurant	Bar	Café	Hotel	Wine Bar	Bakery	Ice Cream Shop	Theater	Beer Bar

Figure 6: Head of data frame showing the 10 most common venue categories extracted around the metro stations representing the neighborhoods.



	Neighborhood	Longitude	Latitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
0	Aksel Møllers Have Station	12.533361	55.686444	2	Café	Coffee Shop	Beer Bar	Bakery	Park	Cocktail Bar	Scandinavian Restaurant	Pub	Wine Bar
1	Amager Strand Station	12.631670	55.656110	1	Beach	Bakery	Coffee Shop	Café	Pizza Place	Gym / Fitness Center	Sushi Restaurant	Burger Joint	Grocery Store
2	Amagerbro Station	12.602944	55.663361	1	Bakery	Coffee Shop	Café	Scandinavian Restaurant	Pizza Place	Bar	Concert Hall	Wine Bar	Gym / Fitness Center
3	Bella Center Station	12.582944	55.638060	4	Restaurant	Supermarket	Other Great Outdoors	Park	Sporting Goods Shop	Convenience Store	Hotel	Café	Discount Store
4	Christianshavn Station	12.591222	55.672220	0	Coffee Shop	Scandinavian Restaurant	Bar	Café	Hotel	Wine Bar	Bakery	Ice Cream Shop	Theater

Figure 7: Head of data frame showing the most common venue categories extracted around the metro stations representing the neighborhoods, including geo-data and cluster label.

4 Results

The following section shows the visualized results of the analysis. Figure 8 shows the coropeth map of the city districts of Copenhagen colored by the current housing price per square meter. It can be seen, that apartments in the city center are most expensive, followed by the area in the south. In the northern districts on the other hand the prices are lower. From this figure an easy overview over affordable areas for the individual buyer can be gained.

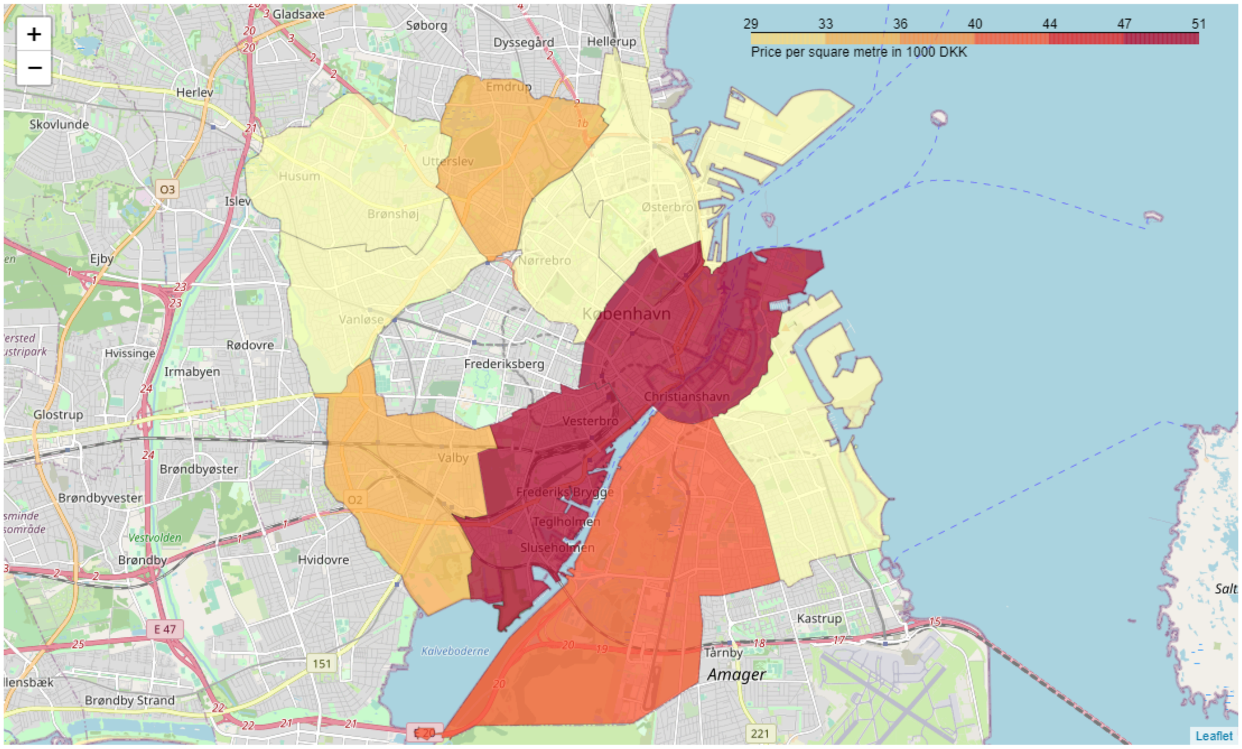


Figure 8: Coropeth map of current housing prices for apartments in Copenhagen.

Figure 9 shows the average annual increase of housing price per square meter for apartments in Copenhagen. The Figure shows that one district in the north (Bispebjerg) had a significantly higher price increase over the last 5 years. Comparing the figure to the previous figure, it can be seen, that the average price for the same district is in the medium range. This shows that interesting opportunities for buying an apartment in Bispebjerg can be expected from a financial point of view.

Figure 10 shows the markers for stations colored by the clusters of similarities. The figure shows the display example for Skjolds Plads Station and the most common venues, namely beer bars, playground, coffee shops, gyms and bakeries. Most clusters separate very clearly in certain areas. The green cluster is dominated by the airport and the airport venues. The purple cluster is a collection of outdoor spaces. The red cluster is characterized by theaters but also tourist venues of the center. Blue is characterized by a mix of bars and coffee shops, but also attractive places for families like playgrounds and gyms. The yellow cluster is characterized mainly by venues for daily living such as grocery stores, schools and playgrounds. This visualization makes it easy to find an attractive area to search for a place to live.

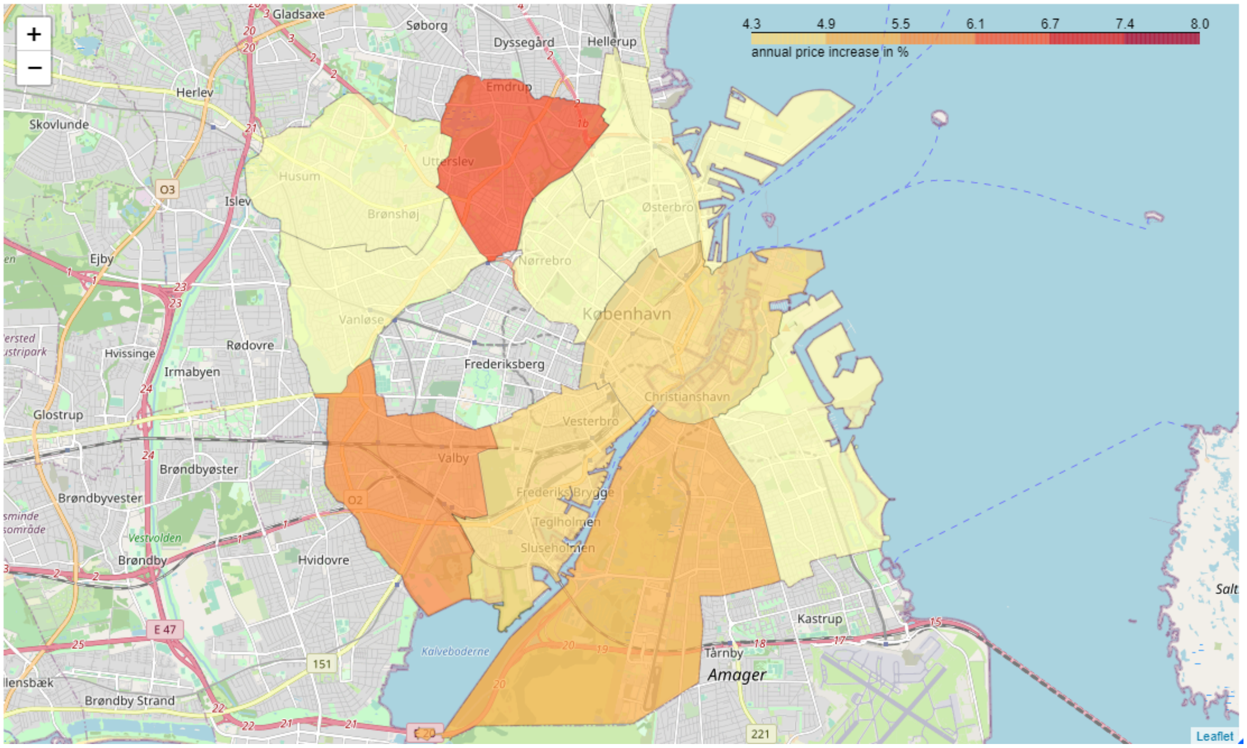


Figure 9: Coropeth map of the average annual increase of housing price per square meter for apartments in Copenhagen.

5 Discussion

The shown visualization of housing prices and neighborhoods in Copenhagen make it easy to find a district of interest for the individual buyer regarding the budget and also the investment perspective. The similarity of neighborhoods helps additionally to decide in more detail which neighborhood is an attractive place to live for the individual person. However, the coropeth maps are not very detailed and the prices of the individual neighborhoods within the districts is not reflected.
Using the metro-network as indicator for neighborhoods has the advantage that it is one of the most important means of daily transportation in Copenhagen. However, it does not cover the full city and leaves opportunities out, especially in the south west of Copenhagen. It is recommended to add the S-train network to the analysis of neighborhoods to reach also the outer parts of Copenhagen and show more opportunities to customers.

6 Conclusion

Over all, this project provides a good help to make initial decisions regarding the district of interest for buying an apartment. It also helps to evaluate offers for homes regarding the price in the neighborhoods as well as the financial perspective of the investment. It can also help to find neighborhoods that are attractive to individual people. It also has to be pointed out, that the project could be improved with more detailed pricing data for the neighborhoods within the districts, and also by including the s-train network to the indicator of neighborhoods.

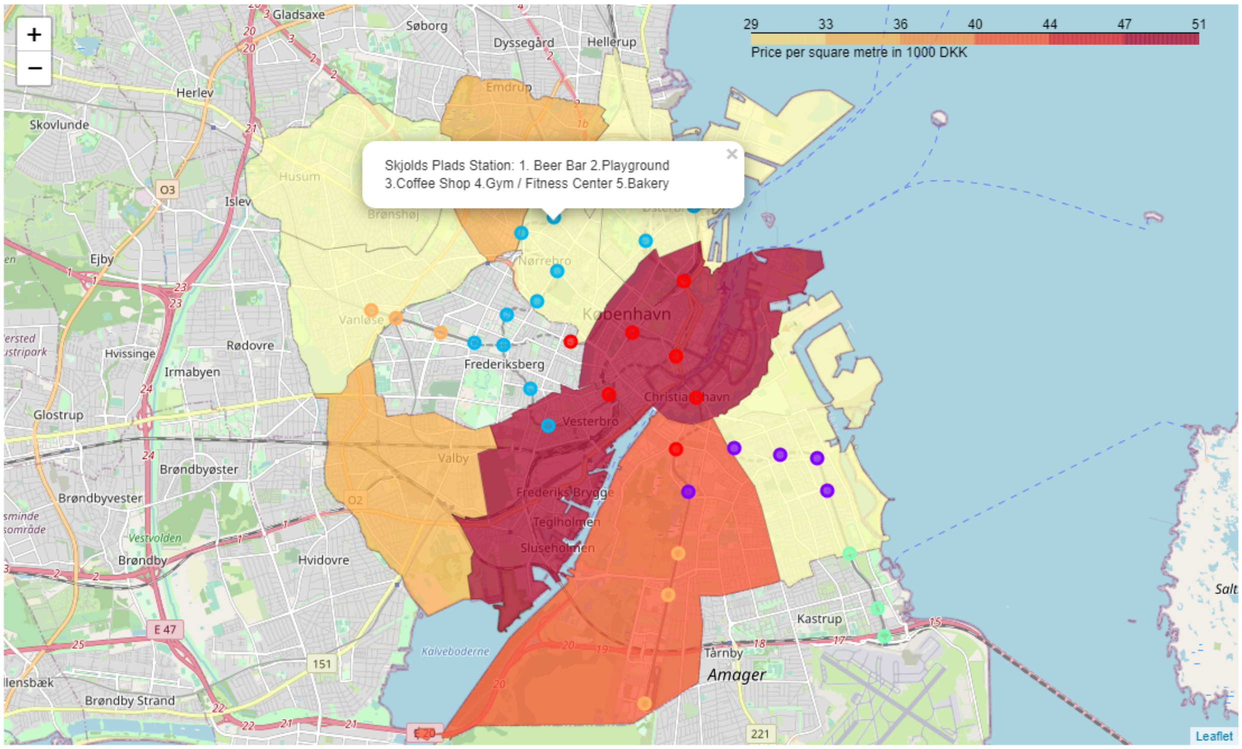


Figure 10: Coropeth map of current housing prices for apartments in Copenhagen and the markers for locations of metro stations clustered by similarity of the neighborhoods.