Applied Data Analysis Capstone Project
Buying a home in Copenhagen

Gesine Wanke

July 11, 2020

1 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.

1.1 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 (((12.5383043143
bydel.4	23	5	Valby	MULTIPOLYGON (((12.5243376363
bydel.5	24	4	Vesterbro-Kongens Enghave	MULTIPOLYGON (((12.5444786733
bydel.6	26	9	Amager st	MULTIPOLYGON (((12.6308225523
bydel.7	21	7	Brnshj-Husum	MULTIPOLYGON (((12.4689396523
bydel.10	22	6	Vanlse	MULTIPOLYGON (((12.4980184363
bydel.8	19	3	Nrrebro	MULTIPOLYGON (((12.5370429343
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.

1.2 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

	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`.

per square meter in 1000 DKK for apartments can be found.

1.3 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

	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`.

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	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.

1.4 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

	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`.

longitude) as well as the venues, their respective geo-data and the category of the venue.