FINAL REPORT – IBM DATA SCIENCE CAPSTONE PROJECT

Challenge: Finding the most suitable investment opportunity for a real estate firm in the province Utrecht in the Netherlands

Author: Arne van den Bosch

MARCH 7, 2021

Table of Contents

1. Background

This notebook has been made as part of the course "IBM Data Science Professional Certificate", which is an introductory course to the field of Data Science. The course consists of 10 modules that cover a variety of aspects of the data science field, including Data Science methodologies, SQL, analysing and visualizing data, machine learning models, etc. As from module 4, the participants start learning the basics of programming in Python. This report is part of a final assignment of the course in which the participants apply the learned theory in practice by coming up with a self-chosen project. As part of the project, geospatial data should be obtained using the Foursquare API. Foursquare is a technology company focused on location data, which is used for apps such as Uber, Apple's Maps, Twitter and a variety of other well known organisations. Via their website, it is for instance possible to look for venues, such as restaurants or shopping facilities in a chosen area and obtain information about those venues, such as customer reviews, photos and a description of the venue itself. Here is an example of a sandwich restaurant in the city of Utrecht in the Netherlands (link).

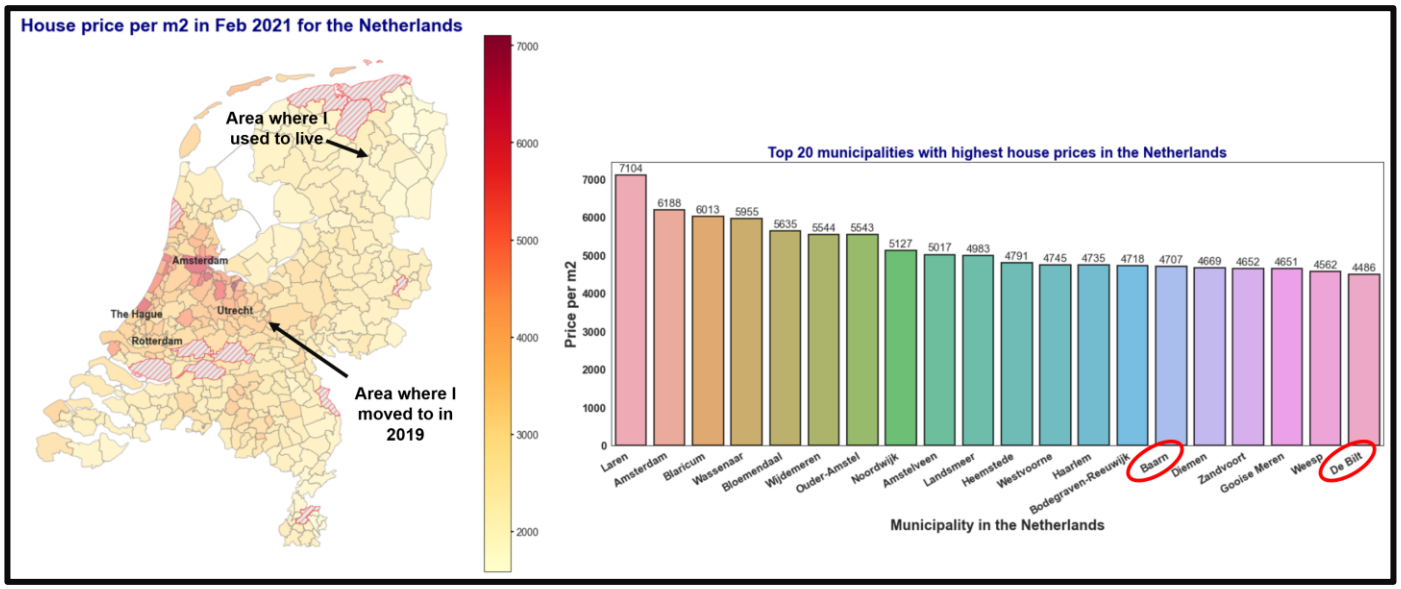
More information about this course can be found on Coursera (link).

2. Introduction / Business Problem

2.1 Prologue

In may 2019 I moved from the upper north of the Netherlands to the province Utrecht, which is one of the 12 provinces in the Netherlands. House prices in the area where I used to live up north of the country were a lot lower than house prices in Utrecht, which is the fourth largest city in the Netherlands. Finding a new place for a reasonable price and with a variety of venues nearby was challenging. Data science skills to gather and to study data would have come in handy during that time. Recently I came across a news article that claimed that house prices in the past 20 years have not risen as fast as in the past quarter. There is a great shortage of houses in the Netherlands, whilst the demand for it keeps rising, driving up the house prices further. Whilst people that like to purchase their first house have great difficulty finding a place that they can afford, it is a lucrative business for professional and private real estate investors, who purchase and resell properties. The situation of moving from a 'cheaper' area to a more 'expensive' area and the challenges that came with finding a place for a reasonable price has given me the inspiration for the self-chosen project for the final assignment of the IBM Data Science course on Coursera. I will further clarify the business problem for the self-chosen challenge in the next paragraph.

Below choropleth map of the Netherlands and the bar chart show the average house prices per m^2 in the Netherlands. The orange/red-coloured area in the west of the Netherlands and in the middle of the country indicate areas with high house prices. The area this report will be about is the province Utrecht, which is indicated on the map below. The highest house prices in Utrecht are in the municipalities Baarn and De Bilt.



2.2 The business problem

For this assignment, let's assume a fictional situation in which I am working for a real estate investment agency. This agency has successfully purchased and resold houses in the larger cities in the Netherlands, such as Amsterdam, Rotterdam and The Hague. However - the agency wants to expand its activities to the fourth largest city in the Netherlands as well, which is Utrecht. The agency gives me the assignment to examine potential interesting areas for purchasing and reselling properties. The term 'interesting area' means: areas that have high potential for a good return on investment, so a good profit will be obtained from it when reselling the property.

I think the specified challenge gives a great opportunity to combine a large variety of freely available data on the internet with the Foursquare geospatial data as required by the assignment. I would like to note that many factors can be taken into consideration for determining the 'sweet spots' in terms of potential interesting areas to invest in properties. For this assignment however, I am going to assume the following factors as being important for potential areas of interest:

- House prices over the past 5 years have a stable increasing trend in the area of interest;
- Ideally the house price is currently reasonably low, but has a historical trend that indicates that house prices in that area will rise further in the coming future.
- There is a variety of venues nearby, such as supermarkets, restaurants, shopping malls, etc.
- Some additional factors that can be beneficial:
  - Low crime rates / nuisance figures;
  - Low energy consumption;
  - Low number of people using employment allowance;
  - High number of organisations nearby (employment opportunities)

3. Description of the data

The data used to find answers to the business problem come from different sources, which are:

1. House prices per square meters per municipality in the Netherlands, as Excel file from the website **"Huizenzoeker.nl"** (link).
2. GEOJSON file with geospatial data for each municipality in the Netherlands from website **"hub.arcgis.com"** (link).
3. Venue names, categories and location coordinates are obtained by using the **Foursquare API**. A description of Foursquare was given at chapter 1, under the heading "Background".
4. Variety of municipality related data (i.e. crime rates, unemployment figures, etc) from the **Dutch Central Bureau of Statistics (CBS for short)** using their **API and python library "cbsodata"** (link).

Data source 1: Huizenzoeker.nl:
This website has an Excel file containing the total average house prices, median house prices, average house prices per square meter and the number of houses for sale. The link for this Excel file is given at point '1' at the top of this chapter. For this assignment we will use the average house price per square meter to study the house price trends over the past years. This average house price per square meters will also be plotted on choropleth maps for each of the 26 municipalities of Utrecht.

The dataset with house-prices was obtained by using the pandas library's "read_excel()" function. Because the dataset was sufficiently complete and clean there was not a lot of pre-processing and data cleansing required in order to make it useful for further analysis.

In context of the business problem this data will help in finding the municipalities that have had increasing and stable house price increases over the past few years.

Data source 2: hub.arcgis.com:
The GEOJSON file is used to obtain the polygon coordinates of the municipalities in the province of Utrecht. Utrecht has 26 municipalities at the moment. In order to plot the shapes of those municipalities on a choropleth map, we will use the 'geopandas' library. This library has a 'gpd.read_file()' function, which allows a GEOJSON file to be retrieved from the internet. The function puts the data in a geopandas-data frame. The data is filtered

for municipalities in the province Utrecht only and is merged with the dataset of Data Source 1. The merged data frame allows house prices per municipality to be visualized in a choropleth map.

Data source 3: Foursquare API:
As requested by the assignment, Foursquare data will be used to obtain venue information, such as location and venue categories (i.e. restaurants, supermarkets, drug store, cafes, monuments, etc) for each of the 26 municipalities. A function will be specified in Python in order to obtain venues in a 5km radius from the centre points of the 26 municipalities. A data frame will be created in order to show the top 10 most common venues. After that, a non-supervised algorithm - the k-means algorithm - will be used to divide the dataset into clusters based on the geographical coordinates of the venues. The clusters are formed based on how near the venues are to each other. The clusters are shown on a folium choropleth map and for each cluster a bar plot is made in order to study the results and examine the distinguishing characteristics of each cluster.

In context of the business problem, the Foursquare venue data will help determine potential popular areas to live. Areas with a variety of venues in the vicinity are generally more popular than places that are remote.

Data source 4: Dutch Central Bureau of Statistics (CBS for short):
When looking for freely available data on the internet about the Netherlands there is a great chance that you will end up on the website "opendata.cbs.nl". CBS was founded in 1899 by the government with the task to collect, process and to publish statistics that would be freely available to society. At the end of the 19th century there was an increase in the amount of social malpractices caused by the large-scale industrialisation process. The government saw the establishment of CBS as a solution to this problem. By providing publicly available statistics social inequalities would be reduced and ministries would use reliable and accurate data. Over the years CBS has grown into a modern, innovative organisation and has a vast database with statistics about a large variety of subjects. More information about the CBS can be found on their website (Link to CBS). Their data is published on CBS Open data StatLine (Link) and can be downloaded in a variety of formats, such as csv- or JSON format. Most of the information they publish is also easily accessible for users of Python and R via their API (Link).

For the assignment I accessed a database containing key figures about municipalities and districts in the Netherlands (Link). I used the API of the CBS and the library "cbsodata" to directly obtain the dataset into a pandas data frame. Then I created a new data frame which contained the columns of interest and renamed the columns to make it more descriptive. I filtered the data frame for the 26 municipalities in province Utrecht and examined the quality and completeness of the data. It turned out that on the level of municipalities, the dataset I obtained had no missing data. I turned some columns with total counts into relative figures by dividing the totals by the number of inhabitants in the municipality and multiplying the result by 100.000, such that the figures for those variables could be compared between different municipalities. The data frame was then merged with the data frame of "Data source 2" mentioned above in order to be able to put the figures on choropleth maps. Finally I plotted the figures in both choropleth maps, complemented by bar charts that show the figures in a descending order.

The key figures I chose from the dataset for further analysis are:
1. Average Electricity Consumption in kWh
2. Average Gas Consumption in m3
3. Average Income per inhabitant x 1000Euro
4. Percentage of households with low income
5. Number of people using welfare allowance
6. Number of people using unemployment allowance
7. Number of thefts from house or barn
8. Number of registered cases of destructive behaviour
9. Number of companies and organisations

Variable number 5, 6, 7 and 8 were turned into relative figures by dividing the amount by the number of inhabitants of the municipality and multiplying the result by 100.000 for the purpose of being able to compare the figures of different municipalities to each other.

Variable number 1 and 2 in the list above may give insight about for instance areas that have newly build homes, which consume less energy or where solar panels, district heating or thermal storage systems may be present, thereby reducing electricity / gas consumption. Energy efficient homes are becoming more and more popular in the Netherlands. Variable 3, 4, 5 and 6 could give us some information about the most dominant

social classes in each municipality, which could be interesting as it could help indicate whether a particular area would be more or less suitable for investing in real estate. For instance, areas where people have high average income and where less people are using an unemployment allowance are preferred over areas that have low income and higher percentages of people using unemployment allowances. Variable 7, 8 and 9 can indicate how save a particular municipality is and variable number 10 may be interesting from the perspective of employment opportunities. The assumption is then that areas with more companies/organisations would also have more job opportunities.

4. Methodology

In this section, I will clarify how data was retrieved from the different sources, what analysis where performed and how conclusions where drawn from the visualisations.

4.1 Part 1: Obtaining and analysing house price figures

In part 1, I obtained the house price data from the website "Huizenzoeker.nl". The data was available in Excel file format from the website and was retrieved using the pandas library and its "read_excel()" function.

In the pre-processing stage I checked the data for missing data and renamed columns to a more descriptive name.

When examining the data it turned out that 3 older municipalities in Utrecht were included in the dataset, which had become one municipality in the year 2019. To prevent problems when merging the dataset with geospatial data later on, the 3 older municipalities were replaced by one municipality. I simply created a new data frame containing the datasets of the 3 municipalities that had to become one. Then I took the average of the house prices of the 3 older municipalities and put the result back in the original data frame. Then I removed the rows of the older 3 municipalities. There were no ‘null-values’ in the data frame, but some rows from municipalities where house prices were not known were filled up with ‘0’ values. This however was not an issue in my case, as the dataset for the 26 municipalities in the province Utrecht that I was interested in was complete. In ended up with the following data frame.

[6]:

	Municipality	Sep 2008	Oct 2008	Nov 2008	Dec 2008	Jan 2009	Feb 2009	Mar 2009	Apr 2009	May 2009	...	May 2020	Jun 2020	Jul 2020	Aug 2020	Sep 2020	Oct 2020	Nov 2020	Dec 2020	Jan 2021	Feb 2021
0	Laren	5336	5272	5208	5120	5062	4990	4980	4985	4969	...	6458	5930	5705	5715	5806	5794	5920	6064	6929	7104
1	Amsterdam	3634	3651	3669	3666	3647	3627	3619	3628	3634	...	5948	6016	6016	6062	6080	6031	6087	6139	6200	6188
2	Blaricum	5711	5582	5495	5255	5021	4942	5093	5278	5321	...	5473	5947	6063	6091	6283	6095	6124	6235	6245	6013
3	Wassenaar	4668	4715	4715	4639	4529	4447	4443	4443	4456	...	5398	5605	5742	5791	5763	5807	5819	5857	6000	5955
4	Bloemendaal	5259	5235	5198	5103	5104	5145	5057	4964	4938	...	5998	5953	6045	6270	6163	6225	6179	5958	5688	5635
5	Wijdemeren	4173	4164	4156	4172	4106	3901	3847	3841	3848	...	5147	5035	5045	5074	5238	5372	5450	5425	5551	5544
6	Ouder-Amstel	3019	3086	3258	3280	3202	3238	3187	3217	3315	...	4935	4926	4838	4859	4879	4982	5127	5110	5377	5543
7	Bergen (NH)	4261	4268	4275	4224	4186	4182	4181	4173	4188	...	5122	5072	5071	5046	5112	5167	5265	5327	5436	5288
8	Noordwijk	3538	3548	3570	3517	3517	3576	3627	3576	3559	...	5181	5136	5091	5001	4955	5069	5151	5078	5132	5127
9	Amstelveen	3108	3118	3170	3137	3110	3104	3078	3043	3018	...	4913	4864	4820	4821	4846	4856	4928	4954	4988	5017

10 rows × 151 columns

Data frame 1: Avg_House_price_per_m2_NL_df

After that, the geospatial data was obtained from the website “opendata.arcgis.com”, which offered the information in geojson-format. The geopandas library and its “read_file()” function was used to retrieve the information from the website and information from the geojson-file was put directly into a geopandas data frame. A new data frame was created in which only the required columns from the geopandas-data frame were selected to end up with the following table.

	Municipality	st_areashape	st_lengthshape	geometry
0	Appingedam	2.462025e+07	25998.736105	POLYGON ((6.80518 53.31672, 6.80510 53.31669, ...
1	Delfzijl	1.356711e+08	87605.267135	POLYGON ((6.78907 53.40235, 6.78721 53.40167, ...

Data frame 2: Geo_Coord_Municipalities_NL_df

In the step thereafter, data frame 1 containing the house price data and data frame 2 containing the geospatial data were merged into one data frame, using the “Municipality column in both data frames as the joining column. In this case a left-join was applied, in which the data from data frame 1 was merged to all the corresponding Municipality columns that were in data frame 2. This yielded the following data frame.

	Municipality	st_areashape	st_lengthshape	geometry	Sep 2008	Oct 2008	Nov 2008	Dec 2008	Jan 2009	Feb 2009	...	May 2020	Jun 2020	Jul 2020
0	Appingedam	2.462025e+07	25998.736105	POLYGON ((6.80518 53.31672, 6.80510 53.31669, ...	1735.0	1717.0	1717.0	1768.0	1799.0	1788.0	...	1833.0	1936.0	1931.0
1	Delfzijl	1.356711e+08	87605.267135	POLYGON ((6.78907 53.40235, 6.78721 53.40167, ...	1640.0	1647.0	1665.0	1658.0	1646.0	1649.0	...	1629.0	1678.0	1670.0
2	Groningen	1.978564e+08	98196.660775	POLYGON ((6.47634 53.23591, 6.47610 53.23537, ...	2015.0	2003.0	1999.0	1999.0	1997.0	2001.0	...	2650.0	2653.0	2591.0
3	Loppersum	1.119941e+08	64460.648494	POLYGON ((6.71516 53.37337, 6.71408 53.37282, ...	1691.0	1707.0	1709.0	1741.0	1739.0	1727.0	...	1804.0	1846.0	1811.0
4	Almere	1.393973e+08	58860.779608	POLYGON ((5.16002 52.39639, 5.15684 52.39295, ...	2009.0	2014.0	2019.0	2023.0	2015.0	2010.0	...	2793.0	2811.0	2854.0

Data frame 3: Geo_df_avg_house_price_per_m2_NL

After merging, another data frame was created in which only the house price per m^2 in February 2021 was shown and an extra column was created in which the average house price increase per m^2 was calculated over the last 6 years. This resulted in the data frame below:

	Municipality	House_price_per_m2_Feb_2021	House_Price_increase_per_m2_Feb_2015_to_Feb_2021	st_areashape	st_lengthshape	geometry
0	Appingedam	2177.0	503.0	2.462025e+07	25998.736105	POLYGON ((6.80518 53.31672, 6.80510 53.31669, ...
1	Delfzijl	1938.0	504.0	1.356711e+08	87605.267135	POLYGON ((6.78907 53.40235, 6.78721 53.40167, ...

Data frame 4: Geo_df_avg_house_price_per_m2_NL_Feb_2021

Then extra columns were added to data frame 4:

- The column “Centre_Point was added by using the geopandas “centroid’ method. This method can be used to calculate the latitude- and longitude for the centre point of the polygons. The result is a POINT variable including the latitude and longitude.
- The column “Long” and “Lat” contain the longitude and latitude of the midpoint of the polygons. The latitude and longitude were extracted from the “Center_point” column, using a “map()” function in combination with a lambda function.
- The column “Coordinates” was added by concatenating the values of columns “Long” and “Lat”. The result was a text object in the column “Coordinates” which was converted to a tuple using “ast” module in python and the function “literal_eval” to turn the text into datatype tuple.

The resulting data frame is shown below:

	Municipality	House_price_per_m2_Feb_2021	House_Price_increase_per_m2_Feb_2015_to_Feb_2021	st_areashape	st_lengthshape	geometry	Center_point	Long	Lat	Coordinates
0	Appingedam	2177.0	503.0	2.462025e+07	25998.736105	POLYGON ((6.80518 53.31672, 6.80510 53.31669, ...	POINT (6.84963 53.31709)	6.849634	53.317086	(6.849633600729762, 53.31708587700029)
1	Delfzijl	1938.0	504.0	1.356711e+08	87605.267135	POLYGON ((6.78907 53.40235, 6.78721 53.40167, ...	POINT (6.92951 53.32011)	6.929508	53.320110	(6.929507506553045, 53.32010952123877)
2	Groningen	2759.0	838.0	1.978564e+08	98196.660775	POLYGON ((6.47634 53.23591, 6.47610 53.23537, ...	POINT (6.62066 53.21863)	6.620659	53.218633	(6.620659225096424, 53.21863257047191)
3	Loppersum	1865.0	317.0	1.119941e+08	64460.648494	POLYGON ((6.71516 53.37337, 6.71408 53.37282, ...	POINT (6.72815 53.34070)	6.728147	53.340698	(6.728147238081251, 53.340697641881924)
4	Almere	3091.0	1156.0	1.393973e+08	58860.779608	POLYGON ((5.16002 52.39639, 5.15684 52.39295, ...	POINT (5.24692 52.36694)	5.246920	52.366940	(5.246919556937503, 52.36693975508572)

Data frame 5: Geo_df_avg_house_price_per_m2_NL_Feb_2021_w_coord

The objective of adding the coordinates to the data frame is to be able to plot labels of municipalities in the choropleth maps that were created and also to collect venue location information within a specific radius from the centre point of each municipality.

Data frame 4 and 5 include the municipalities for the entire country ‘The Netherlands”. Similar data frames were created for the province Utrecht. A list was made in Python containing the names of the 26 municipalities in Utrecht. Then the pandas “isin()” function was applied to create data frame 6 and data frame 7 below, with a selection of the 26 municipalities that are in the province Utrecht.

Create a list of municipalities in Utrecht. The list will be used to filter the dataframes for municipalities in Utrecht

```
# Creating a list of the 26 municipalities in Utrecht
List_Municipalities_in_Utrecht_Province = ['Amersfoort', 'Baarn', 'Bunnik', 'Bunschoten', 'De Bilt', 'De Ronde Venen', 'Eemnes',
                                           'Houten', 'IJsselstein', 'Leusden', 'Lopik', 'Montfoort',
                                           'Nieuwegein', 'Oudewater', 'Renswoude', 'Rhenen', 'Soest',
                                           'Stichtse Vecht', 'Utrecht', 'Utrechtse Heuvelrug', 'Veenendaal',
                                           'Wijk bij Duurstede', 'Woerden', 'Woudenberg', 'Zeist', 'Vijfheerenlanden']
```

Making a list containing the names of 26 municipalities in Utrecht

```
# Creating a new dataframe by copying all rows with municipalities in Utrecht using the list with municialities in Utrecht.
Geo_df_avg_house_price_per_m2_Utrecht_df_Feb_2021 = Geo_df_avg_house_price_per_m2_NL_Feb_2021[Geo_df_avg_house_price_per_m2_NL_Feb_2021['Municipality'].isin(List_Municipalities_in_Utrecht_Province)].copy()

# Sort the dataframe in descending order based on the last column of the dataframe
Geo_df_avg_house_price_per_m2_Utrecht_df_Feb_2021.sort_values(by=['House_price_per_m2_Feb_2021'], ascending = False, inplace = True)

# Drop and reset the index numbering
Geo_df_avg_house_price_per_m2_Utrecht_df_Feb_2021.reset_index(drop = True, inplace = True)

# Show the dataframe
Geo_df_avg_house_price_per_m2_Utrecht_df_Feb_2021.head(2)
```

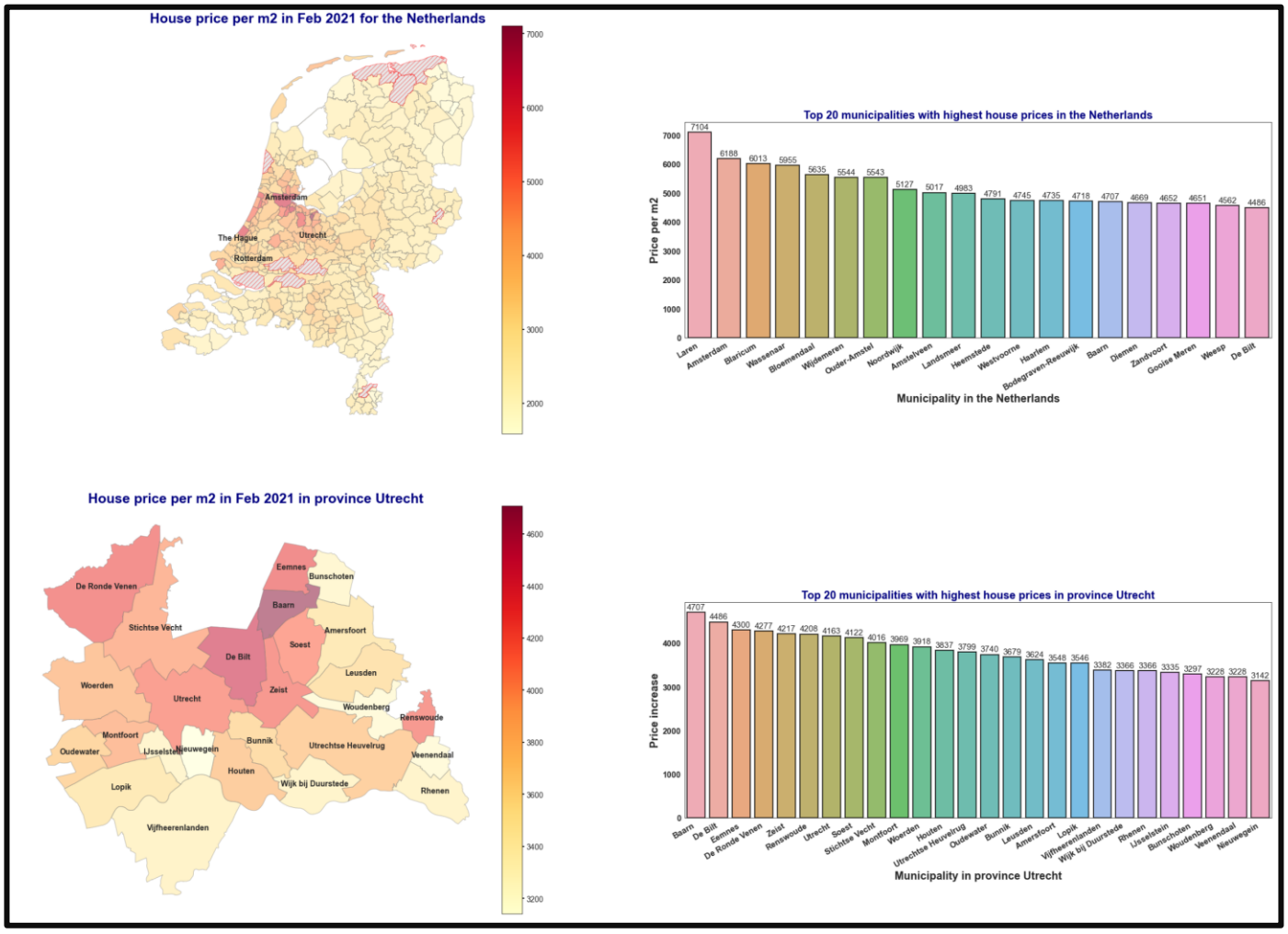
	Municipality	House_price_per_m2_Feb_2021	House_Price_increase_per_m2_Feb_2015_to_Feb_2021	st_areashape	st_lengthshape	geometry
0	Baarn	4707.0	1649.0	3.302635e+07	31522.964656	POLYGON ((5.25588 52.22445, 5.25410 52.22448, ...
1	De Bilt	4486.0	1165.0	6.710147e+07	43218.826211	POLYGON ((5.11861 52.17244, 5.11867 52.17230, ...

Data frame 6: Geo_df_avg_house_price_per_m2_Utrecht_df_Feb_2021.

	Municipality	House_price_per_m2_Feb_2021	House_Price_increase_per_m2_Feb_2015_to_Feb_2021	st_areashape	st_lengthshape	geometry	Center_point	Long	Lat	Coordinates
0	Baarn	4707.0	1649.0	3.302635e+07	31522.964656	POLYGON ((5.25588 52.22445, 5.25410 52.22448, ...	POINT (5.26417 52.20385)	5.264172	52.203848	(5.264172266197837, 52.203847837423176)
1	De Bilt	4486.0	1165.0	6.710147e+07	43218.826211	POLYGON ((5.11861 52.17244, 5.11867 52.17230, ...	POINT (5.17433 52.14173)	5.174335	52.141734	(5.174334940770484, 52.14173369932546)

Data frame 7: Geo_df_avg_house_price_per_m2_Utrecht_df_Feb_2021_w_coord.

In the next step I used the “subplot2grid” functionality from the ‘matplotlib’ library to create 4 subplots at once in one overview. The result is as follows:



Observations from the above subplots are:

The choropleth map of the Netherlands on the upper left hand side clearly indicates that Utrecht is one of the areas with the highest average house prices in the Netherlands. From the bar chart on the upper right hand side it is clear that municipalities in Utrecht are not in the top 10 when considering all the municipalities in the Netherlands. The highest house prices for Utrecht are in municipality Baarn, followed by the Bilt. On the list of places with highest house prices in the Netherlands, Baarn is on place 15 and De Bilt is on place 20.

After creating the subplots, the data from ‘data frame 1’ was used to create a new data frame, including only the average house prices per m^2 for the 26 municipalities in Utrecht. A list of municipalities in province Utrecht was used to select the right municipalities.

```
# Creating a list of the 26 municipalities in Utrecht
List_Municipalities_in_Utrecht_Province = ['Amersfoort', 'Baarn', 'Bunnik', 'Bunschoten', 'De Bilt', 'De Ronde Venen', 'Eemnes',
                                           'Houten', 'IJsselstein', 'Leusden', 'Lopik', 'Montfoort',
                                           'Nieuwegein', 'Oudewater', 'Renswoude', 'Rhenen', 'Soest',
                                           'Stichtse Vecht', 'Utrecht', 'Utrechtse Heuvelrug', 'Veenendaal',
                                           'Wijk bij Duurstede', 'Woerden', 'Woudenberg', 'Zeist', 'Vijfheerenlanden']

# Creating a new dataframe by copying all rows with municipalities in Utrecht using the list with municialities in Utrecht.
Avg_House_price_per_m2_Utrecht_df = Avg_House_price_per_m2_NL_df[Avg_House_price_per_m2_NL_df['Municipality'].isin(List_Municipalities_in_Utrecht_Province)].copy()
Avg_House_price_per_m2_Utrecht_df.reset_index(drop = True, inplace = True)
Avg_House_price_per_m2_Utrecht_df.head(2)
```

	Municipality	Sep 2008	Oct 2008	Nov 2008	Dec 2008	Jan 2009	Feb 2009	Mar 2009	Apr 2009	May 2009	...	May 2020	Jun 2020	Jul 2020	Aug 2020	Sep 2020	Oct 2020	Nov 2020	Dec 2020	Jan 2021	Feb 2021
0	Baarn	3448	3436	3443	3459	3443	3403	3395	3344	3323	...	4455	4436	4307	4253	4301	4562	4670	4483	4659	4707
1	De Bilt	3450	3504	3490	3475	3449	3398	3397	3394	3369	...	4380	4340	4339	4360	4430	4453	4411	4401	4469	4486

Data frame 8: Avg_House_price_per_m2_Utrecht_df

Data frame 8 on previous page was transposed for the purpose of plotting line charts as subplots, which resulted in data frame 9 below. The data frame below was used to loop through in order to make linechart-subplots showing the house-price trends over the years for each municipality.

	Month and year	Baarn	De Bilt	Eemnes	De Ronde Venen	Zeist	Renswoude	Utrecht	Soest	Stichtse Vecht	...	Amersfoort	Lopik	Vijfheerenlanden	Wijk bij Duurstede	Rhenen	IJsselstein	Bunschoten	Veenendaal	Woudenberg	Nieuwegein
0	Sep 2008	3448	3450	3347	3475	3353	3048	2837	3147	0	...	2511	2690	2727	2818	2956	2847	2698	2650	2714	2274
1	Oct 2008	3436	3504	3389	3503	3399	2979	2839	3148	0	...	2520	2726	2777	2901	3001	2811	2628	2658	2771	2289
2	Nov 2008	3443	3490	3312	3502	3382	3026	2863	3119	0	...	2528	2762	2809	2933	3069	2780	2517	2643	2781	2268
3	Dec 2008	3459	3475	3194	3485	3369	3003	2866	3126	0	...	2529	2801	2828	2970	3052	2835	2531	2633	2779	2272
4	Jan 2009	3443	3449	3200	3524	3357	2998	2873	3096	0	...	2519	2808	2770	3015	3049	2820	2565	2626	2780	2289

5 rows × 27 columns

Data frame 9: Avg_House_price_per_m2_Utrecht_df_transposed

Then a copy of data frame 6 was made and the data frame was sorted in descending order for the house price increase per m^2. This yielded the following data frame.

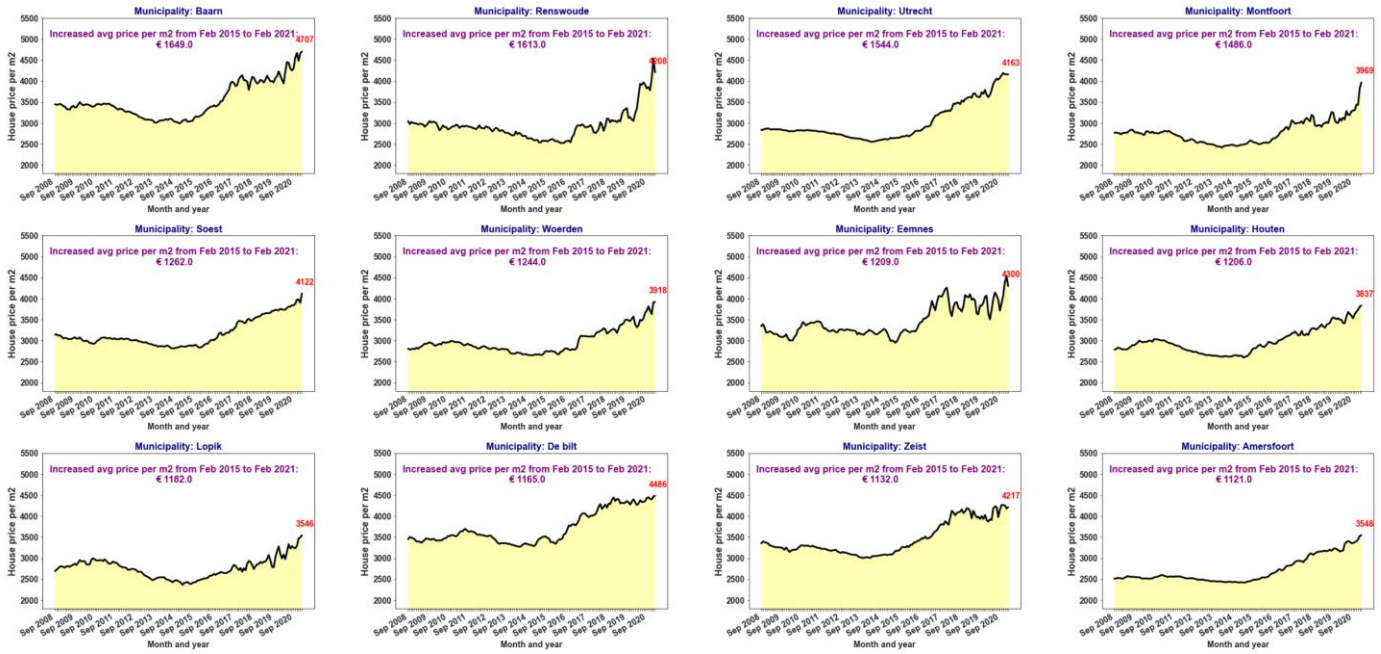
	Municipality	House_price_per_m2_Feb_2021	House_Price_increase_per_m2_Feb_2015_to_Feb_2021	st_areashape	st_lengthshape	geometry
0	Baarn	4707.0	1649.0	3.302635e+07	31522.964656	POLYGON ((5.25588 52.22445, 5.25410 52.22448, ...
1	Renswoude	4208.0	1613.0	1.851735e+07	23947.759304	POLYGON ((5.53150 52.10226, 5.53152 52.10223, ...
2	Utrecht	4163.0	1544.0	9.921474e+07	59611.274509	POLYGON ((4.97864 52.11300, 4.97873 52.11296, ...
3	Montfoort	3969.0	1486.0	3.822334e+07	36460.470124	POLYGON ((4.95682 52.06607, 4.95665 52.06759, ...

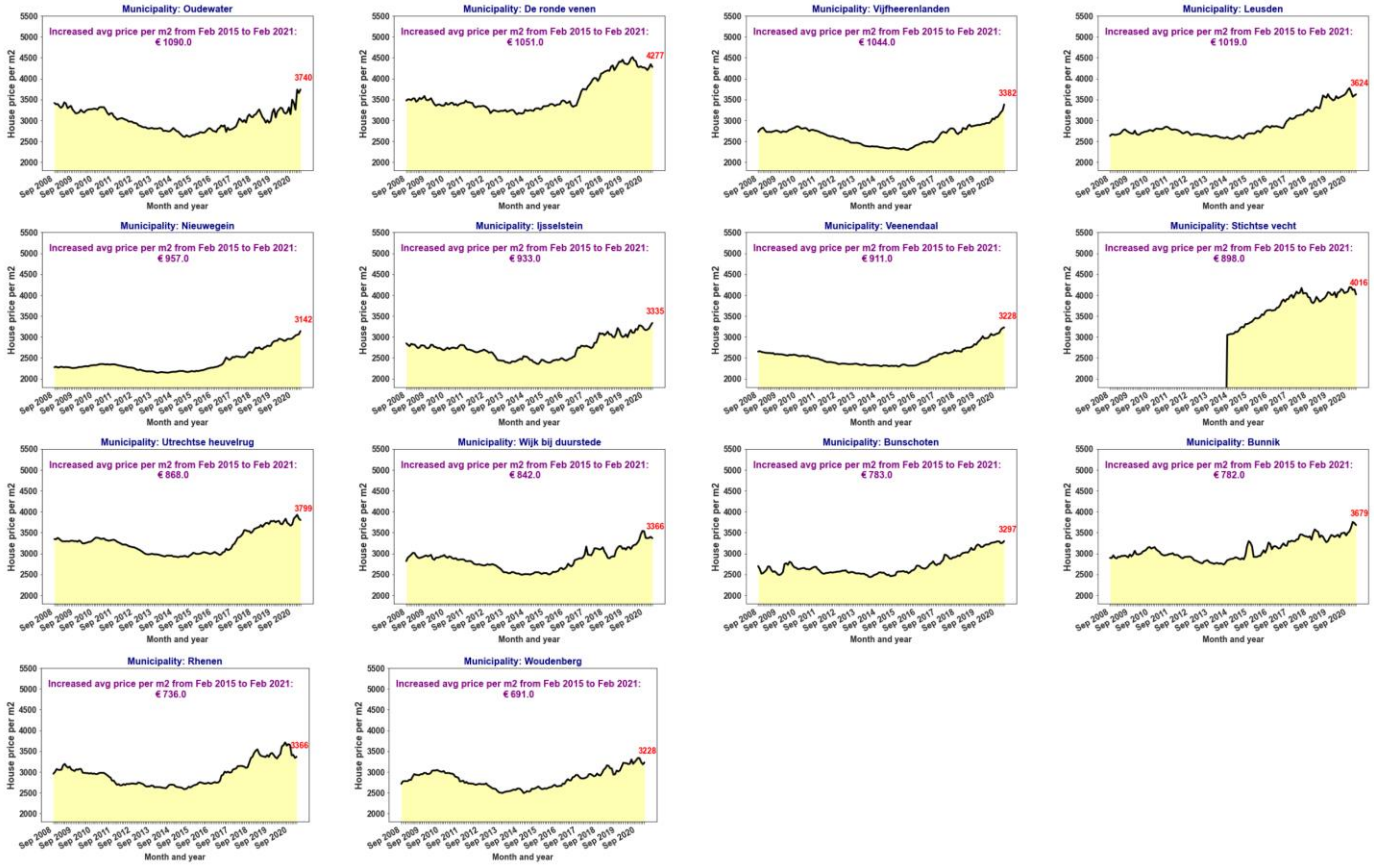
Data frame 10: House_price_per_m2_and_increase

Data frame 10 is used to put a label in the line-charts that indicates the price increase in the last 6 years. It is also used to create a list which is in descending order of house price increase, so the first plot shown is the municipality with the highest house price increase and the last plot shown is the municipality with the smallest house price increase.

In the following step, subplots were generated, using Data frame 9 and Data frame 10, which showed the trends for house price per m^2 over the years for each municipality. A “for loop” was used to go through the list of municipalities and plot a line-chart for each municipality by using the seaborn library. The line-charts were in descending order based on the column with house price increase per m^2 over the past 6 years.

The result was as follows:





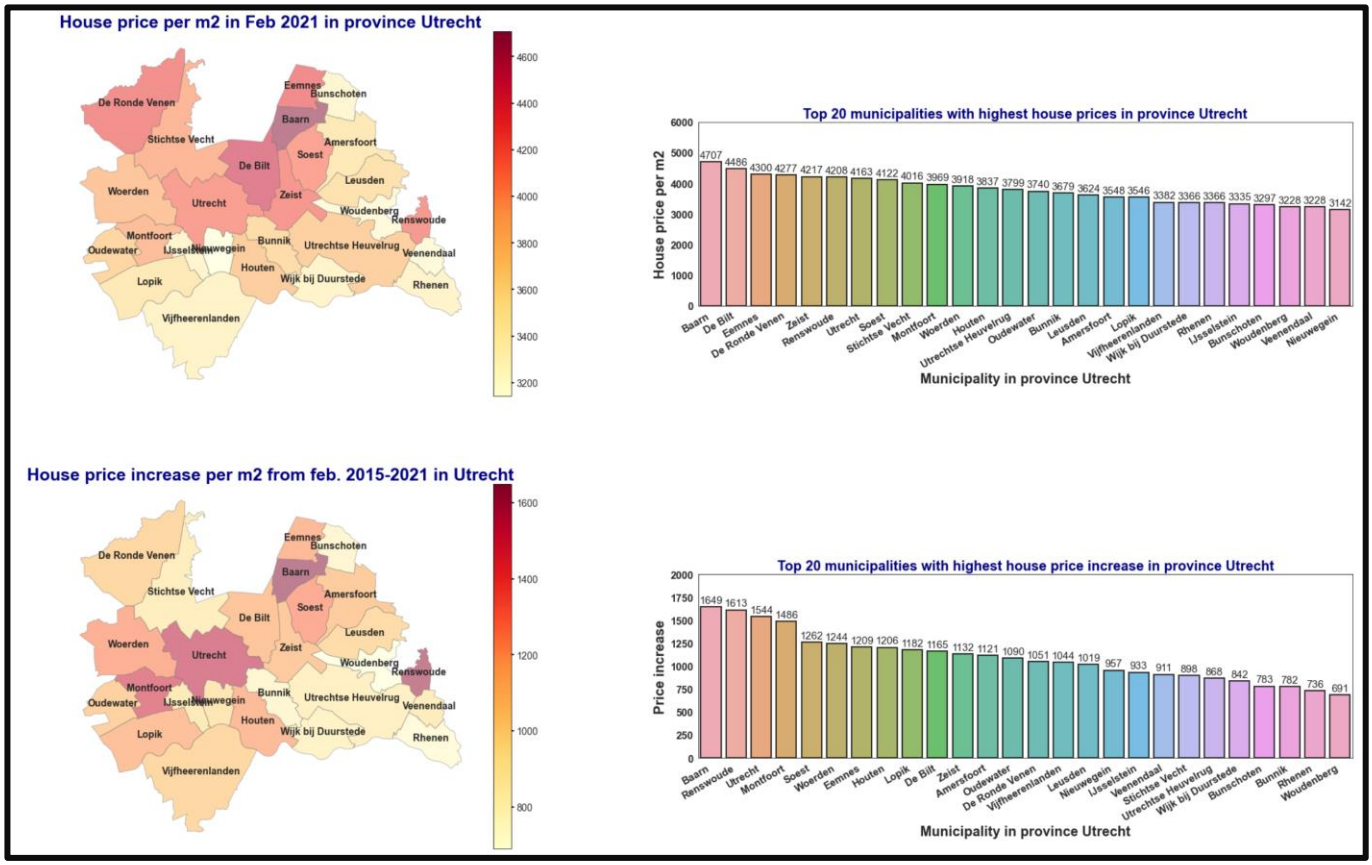
Observations from the above line charts:

From above line charts it is clear that a few municipalities had a steep increase in house price over the past 6 years. Among the top 10 municipalities with highest house price increase are:

- Baarn - €1649 per m^2 house price increase
- Renswoude - €1613 per m^2 house price increase
- Utrecht - €1544 per m^2 house price increase
- Montfoort - €1486 per m^2 house price increase
- Soest - €1262 per m^2 house price increase
- Woerden - €1244 per m^2 house price increase
- Eemnes - €1209 per m^2 house price increase
- Houten - €1206 per m^2 house price increase
- Lopik - €1182 per m^2 house price increase
- De Bilt - €1165 per m^2 house price increase

Municipality Renswoude has had a remarkable increase in house price. In just over 2 years the price increased from about €3.200 per m^2 to €4.208 per m^2. At the moment though the price seems to have a decreasing trend.

Then using the “subplot2grid” functionality from the matplotlib-libary, using the “plot()” functionality from geopandas and using the seaborn library, the following choropleth maps and bar charts were created from data frame 6, 7 and 10.



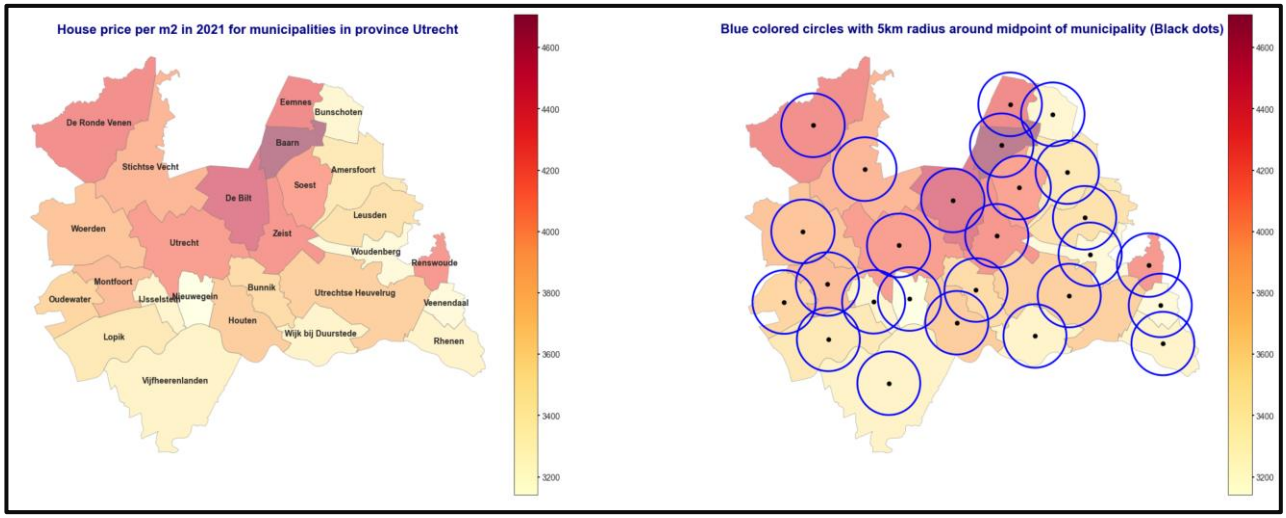
Observations:
Looking at the choropleth maps, a combination of low average house price with high house price increase would be interesting. This means a lighter color in the upper choropleth and darker color in lower choropleth. By looking at the choropleths that way, then the following municipalities seem to be interesting choices:

- **Utrecht:** house prices are rather high, but return on investment seems to be good as house prices in this area have risen a lot during the past 6 years. In fact amongst the 26 municipalities, municipality Utrecht is the third municipality with highest average house price increase over the past 6 years.
- **Nieuwegein:** House prices in Nieuwegein are relatively low, but house prices in this area have had a stable increasing trend over the past 6 years. The municipality is surrounded by other municipalities with higher house prices. As a result, house prices in this municipality may rise as well. Some more research should be done to establish why house prises in this area are currently significantly lower.
- **Amersfoort:** House prices in Amersfoort are of average height and house prices have had a stable increasing trend in the past 6 years.
- **Montfoort:** House prices in Montfoort are also of average height. This municipality has very recently had a very steep price increase.

4.2 Part 2: Collecting and visualizing venue location data

In previous section - part 1 - we only focused on house prices and house price increase. In part 2, we will obtain venue data and use the folium library to plot venues on top of a choropleth map. A non-supervised machine learning technique – K-means - will be applied to cluster venue data based on the geospatial data available.

Prior to using the Foursquare API to obtain venue data the following map was created, including circles around the centre point of each municipality to see what radius around each centre point should be taken when specifying a radius in the syntax to obtain venue data using the Foursquare API.



Observation from the map above:

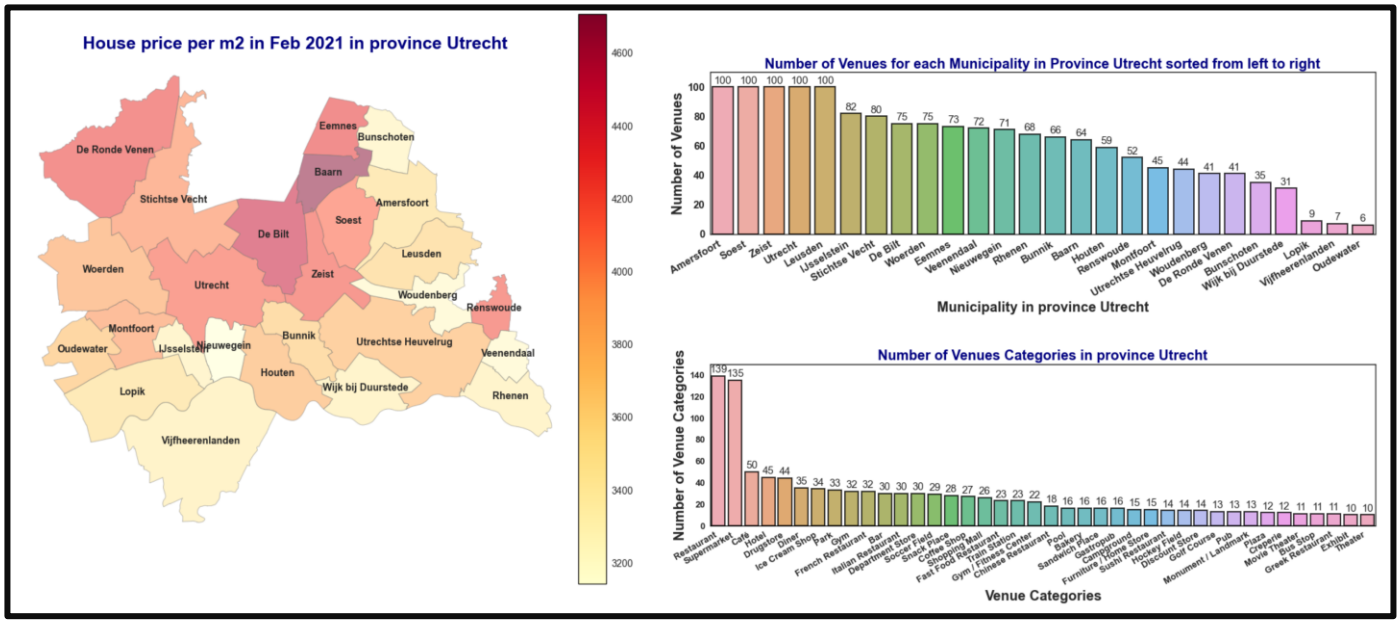
Because area-sizes for the municipalities seem to vary a lot, it is rather difficult to find a radius that suits all areas sizes. I have chosen a radius of 5km. With this radius, it is clear from the map on the right hand side in above picture that some areas have overlapping circles. Hence when looking for nearby venues, there will be some overlap and the same venues will be counted for multiple municipalities. It is also clear from the map that some areas do not fall within a radius and hence some venues that may exist are not part of the analysis. I have searched the internet to apply the polygons of the municipalities as the boundary for searching venues using the Foursquare API. However - I could not find a solution that worked for me and had to accept that a radius would have to be specified.

To obtain venue data for a 5km radius around the centre point of each municipality I started with specifying a function that could retrieve all the required venue information for each municipality at once. The function contained an API Request and Get Request to obtain venue information for the given municipality names, centre point latitudes, longitudes and the given radius and limit for the amount of venues that could be retrieved per search request. The limit for amount of venues per municipality was set to 100 venues and the maximum radius per municipality centre point was set to 5km. After specifying the function, the function was applied to data frame 7. This resulted in the data frame below, including the venue name, venue category and the coordinates in latitude and longitude for each municipality.

	Municipality	Municipality Latitude	Municipality Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Baarn	52.203848	5.264172	Kasteel De Hooge Vuursche	52.203667	5.245899	Monument / Landmark
1	Baarn	52.203848	5.264172	Paleis Soestdijk	52.194061	5.280125	Palace
2	Baarn	52.203848	5.264172	Kasteel Groeneveld	52.218676	5.255431	Monument / Landmark
3	Baarn	52.203848	5.264172	Hoeve Ravenstein	52.216226	5.255728	Farm
4	Baarn	52.203848	5.264172	Vuursche Bos	52.203181	5.234696	Trail

Data frame 11: Nearby_Venues_Municipalities_Utrecht_df

Then subplots were created using data frame 7 and data frame 11, which resulted in the following choropleth map and bar charts.



Observations:

- For the first 5 municipalities the limit of 100 venues found by the Fourquare API has been exceeded and hence the number of venues stop at 100.
- Unsurprisingly, Utrecht, Soest and Zeist are on top of the list in terms of venue count. Those municipalities are in the urban part of the province Utrecht, where average house prices are also in the higher spectrum.
- Amersfoort is also a city and municipality in the province Utrecht. House prices in province Amersfoort are of average height (€3,548 per m2 in February 2021), however the venue count exceeds the 100 counts.
- A similar conclusion can be drawn for the municipality Leusden, the municipality count exceeded 100, but the house prices in Leusden are relatively low (€3624 per m2).
- Other municipalities where house prices are relatively low, but where venue count is relatively high are: Ijsselstein, Nieuwegein, Houten, Veenendaal and Rhenen.

From data frame 11 a selection was made of columns, including the Venue name, Venue Category and location data of the venues. Some of the radiuses overlapped each other, which resulted in venues being counted twice for different municipalities. In data frame 12, those duplicates have been removed.

Create the dataframe that will be used for clustering later on, in which the duplicates are removed that were established in previous step

```
Cluster_df = Nearby_Venues_df.drop_duplicates(subset = 'Check_duplicates_Column').reset_index(drop = True).drop('Check_duplicates_Column',axis=1)
Cluster_df
```

	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Kasteel De Hooge Vuursche	52.203667	5.245899	Monument / Landmark
1	Paleis Soestdijk	52.194061	5.280125	Palace
2	Kasteel Groeneveld	52.218676	5.255431	Monument / Landmark
3	Hoeve Ravenstein	52.216226	5.255728	Farm
4	Vuursche Bos	52.203181	5.234696	Trail

Data frame 12: Cluster_df

In order to apply the non-supervised k-means algorithm, categorical variables had to be dropped from Data frame 12. This resulted in the following data frame.

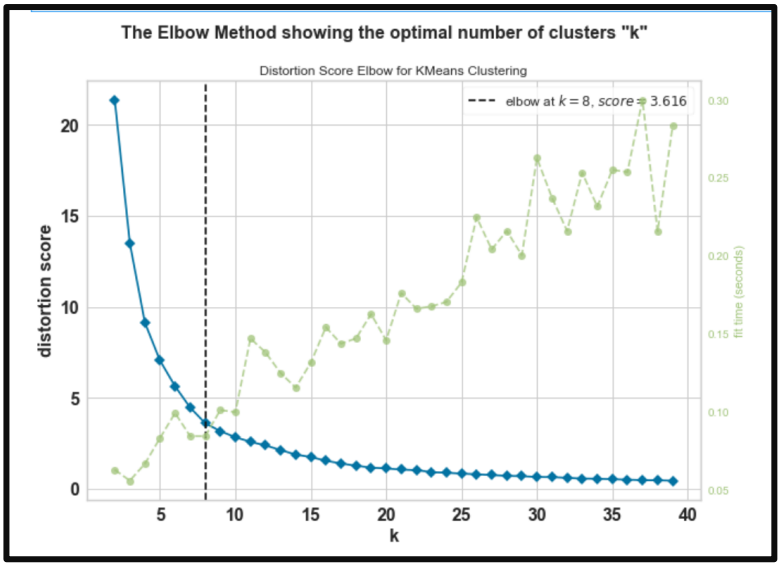
Dropping columns that are not suitable/needed for the k-means algorithm

```
K_Means_df = Cluster_df.drop(['Venue','Venue Category'],axis=1)
K_Means_df.head(4)
```

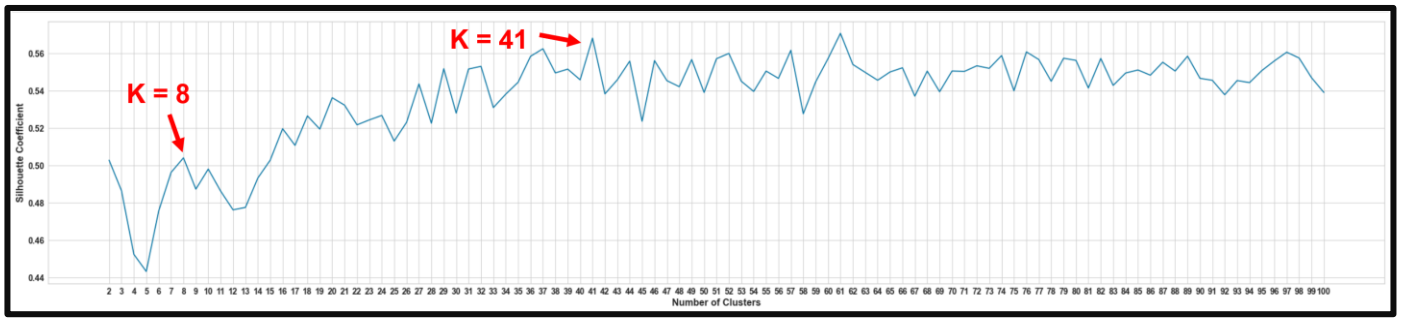
	Venue Latitude	Venue Longitude
0	52.203667	5.245899
1	52.194061	5.280125
2	52.218676	5.255431
3	52.216226	5.255728

Data frame 13: K_Means_df

Then the ‘elbow’ method was used in which a distortion score was plotted against the number of clusters ‘k’. The yellowbrick library was used to create the elbow chart, in which the most suitable amount of clusters ‘k’ was automatically shown by a black coloured vertical dashed line. The most suitable amount of clusters was shown to be ‘8’.



Additionally to the elbow chart, another method was applied, which is referred to as the “Silhouette Coefficient” method. The Scykit-Learn library was used to create below line chart, where the number of clusters ‘k’ is plotted against a calculated Silhouette Coefficient. In general, the higher the silhouette coefficient, the better the cluster is. In below chart I initially thought cluster 41 would be the best choice for the amount of clusters. However – after opening my notebook and after re-running al the code, the graph changed, except for the section at the beginning, including k = 8. This is one of the difficulties I faced during the assignment eventually decided to choose k = 8 for the amount of clusters.



After choosing k = 8, an array was created with cluster labels, using data frame 13 as input.

Create an array with cluster labels using the found optimum value for "K".

```
# Set the number of clusters
K = 8

# Run k-means clustering on the created dataframe
kmeans = KMeans(n_clusters = K,random_state=0).fit(K_Means_df)

# Check cluster labels generated for each row in the dataframe
labels = kmeans.labels_
labels
```

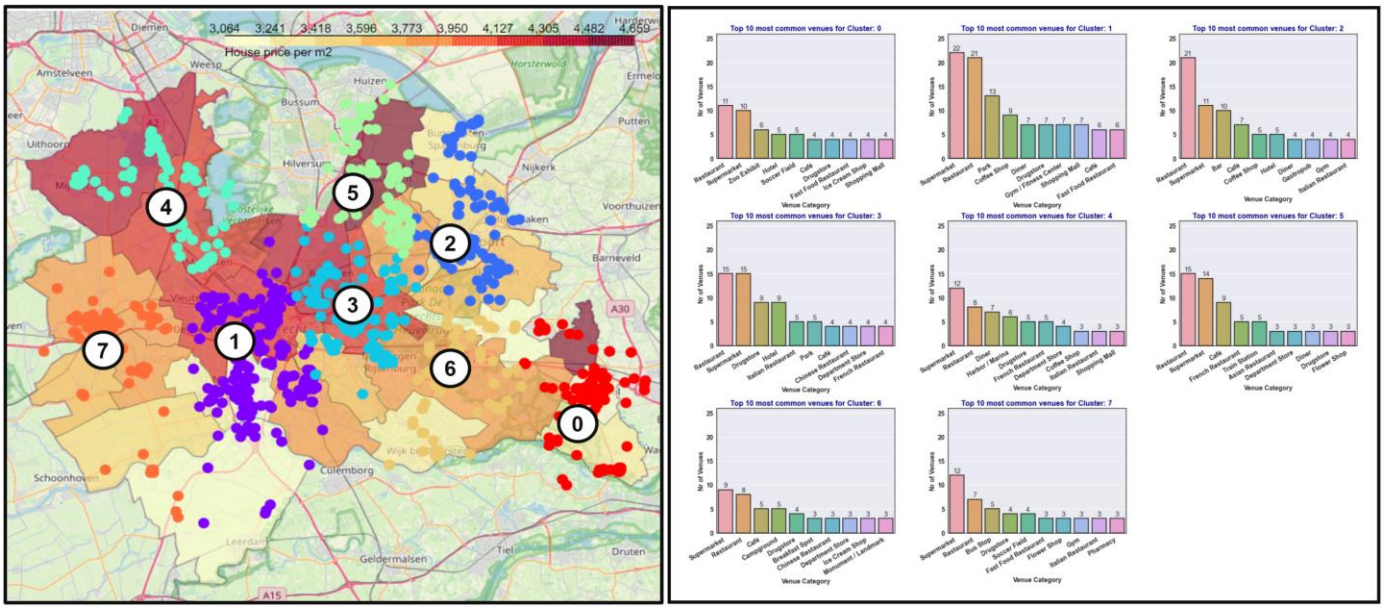
```
array([7, 7, 7, ..., 3, 3, 3])
```

The array with cluster labels was then inserted into the data frame 12, which resulted in the following data frame including the cluster labels.

	Cluster Labels	Venue	Venue Latitude	Venue Longitude	Venue Category
0	0	Ribhouse Texas	52.081564	5.421079	Steakhouse
1	0	Kruidvat	52.010260	5.431644	Drugstore
2	0	Boscafe Sandenburg	52.031625	5.369547	Café
3	0	Restaurant Landgoed Zonheuvel	52.049476	5.353740	Restaurant
4	0	Restaurant Darthuizen	52.015703	5.416660	Chinese Restaurant

Data frame 14: Cluster_df2

In the step after that, a map was created using the folium library. Data frame 6 and 7 were used as input to show house prices in a choropleth map. Data frame 14 was used to show the venues as markers on top of the choropleth map. The cluster labels were visible by the 8 different colours of the markers on the map. Supplementary to the choropleth map, bar charts were made for each of the 8 clusters showing the top 10 most common venues for each of the clusters. A picture of the folium map and bar charts is shown below.



Observations from the choropleth maps and bar charts on previous page:

Overall there seem to be multiple clusters and sub areas in those cluster where it could be interesting to invest in real-estate with the aim of reselling it for a higher price. Cluster 1, 3, 4 and 5 are in the highest house price range and have a high degree of urbanity, with dense areas of venues being closely packed together. Around those clusters there are some areas that also have a high venue count and density, but where house prices are lower, which is the 'south part of cluster 1' and 'cluster 2'. Assuming that we aim for areas with low house prices and potential high profit in the future, then the “south part of cluster 1” and “cluster 2” would be interesting areas to examine further.

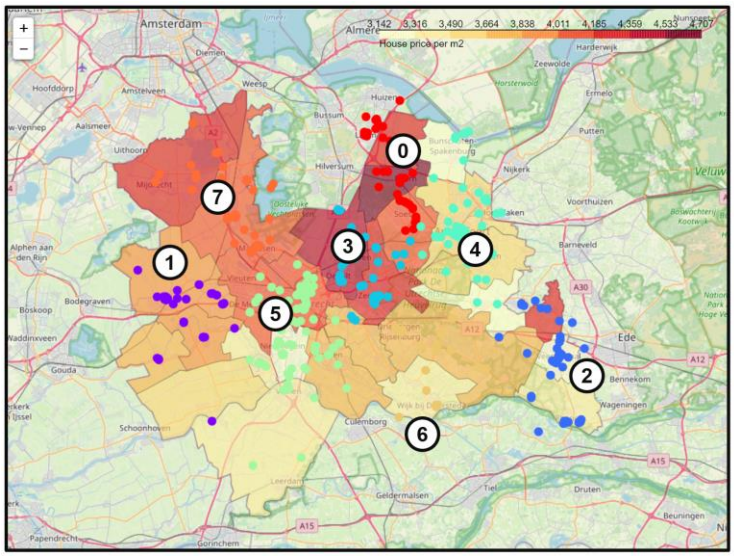
A similar process was repeated for creating a choropleth map with markers on top, but this time only venues for outdoor eating- and drinking facilities were included.

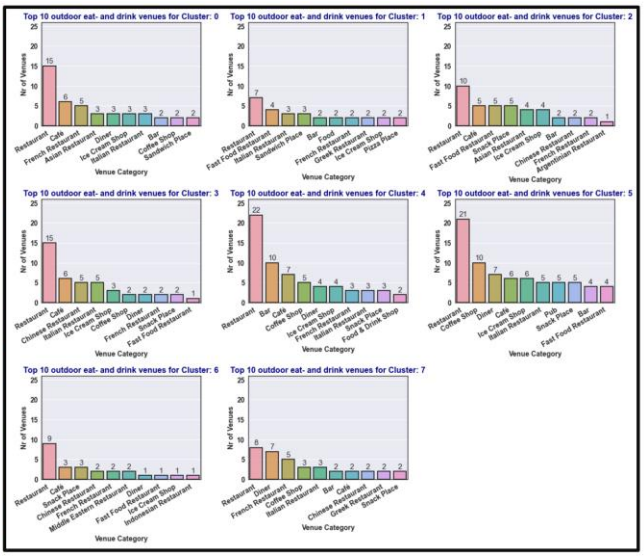
Make a list of venues related to drinking and eating venues outdoor and use the list to select relevant venues from above dataframe

```
List_of_Venues_Eat_or_Drink = ['Restaurant','Café','bistro','brasserie','joint','Cafetaria','Coffee','Bar','Diner','Food','Ice Cream','Pizza','Pub','Salad','Sandwich','Snack','Wine','Drink']
Nearby_Eat_and_Drink_Venues_Utrecht = Cluster_df[Cluster_df['Venue Category'].str.contains('|'.join(List_of_Venues_Eat_or_Drink))].reset_index(drop = True)
Nearby_Eat_and_Drink_Venues_Utrecht.head(5)
```

	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Smaecken van Hamelink	52.211656	5.291049	Ice Cream Shop
1	Restaurant Vuur	52.202181	5.247596	Restaurant
2	The Golden Coffee Box (Boot koffie)	52.211594	5.288511	Coffee Shop
3	IJssalon Bemer	52.187868	5.283528	Ice Cream Shop
4	Cosa cucina & wine bar	52.211967	5.284038	Italian Restaurant

After applying the elbow-method, the k-means algorithm and inserting the cluster labels into the data frame with venue names and categories, the following choropleth map and bar plots were created. As can be seen, the same amount of clusters was found and the clusters look very similar to the one shown on previous page. One big difference is that cluster number 6 on previous page has become very small and exists only of 4 venues in the choropleth map below.





Observations from above choropleth map and bar charts on the right hand side:

- Most dense areas with outdoor eating- and drinking venues are in clusters 4 and 5, followed by clusters 0, 3, 2 and 7. Clusters 1 and 6 have the smallest amount of outdoor eating- and drinking facilities.
- At the south part of cluster 5, house prices are relatively low in municipalities Nieuwegein and Houten but there seem to be a lot of outdoor eating- and drinking venues.
- Also in cluster 4, with the municipalities Amersfoort and Leusden, house prices are relatively low and there are several eating- and drinking facilities.

4.3 Part 3: Visualizing some other characteristics

Some additional information was gathered from the Dutch Central Bureau of Statistics (CBS for short) about municipalities, such as number of thefts, electricity consumption, gas consumption, number of people using unemployment allowance, etc.

The first step was to import the ‘cbsodata’ library in order to use the API of the CBS to directly obtain data in a pandas data frame.

Obtainning data from the website of the Dutch Central Bureau of Statistics (CBS for short), by using their API

```
import cbsodata # CBS library to be able to retrieve data from their website, by just referring to a table number, such as '84286NED'.
District_Data = pd.DataFrame(cbsodata.get_data('84286NED', catalog_url='opendata.cbs.nl'))  # dataderden.cbs.nl    opendata.cbs.nl
District_Data.head()
```

	ID	WijkenEnBuurten	Gemeentenaam_1	SoortRegio_2	Codering_3	IndelingswijzigingWijkenEnBuurten_4	AantalInwoners_5	Mannen_6	Vrouwen_7	k_0Tot1
0	0	Nederland	Nederland	Land	NL00	.	17181084	8527041	8654043	
1	1	Aa en Hunze	Aa en Hunze	Gemeente	GM1680	.	25390	12666	12724	
2	2	Wijk 00 Annen	Aa en Hunze	Wijk	WK168000	1	3560	1735	1825	
3	3	Annen	Aa en Hunze	Buurt	BU16800000	1	3415	1660	1755	
4	4	Verspreide huizen Annen	Aa en Hunze	Buurt	BU16800009	1	145	75	70	

Data frame 15: District_Data

Then a new data frame was made in which the columns of interest were selected from data frame 15. The columns were renamed in order to make them more descriptive. The data of type object (text data) turned out to have leading and trailing white spaces, which were removed by applying the ‘strip()” function. The data frame was filtered using a list with names of the 26 municipalities in Utrecht. These pre-processing actions resulted in the following data frame.

	Municipality	Nr of inhabitants	Avg_Electr_Consump_kWh	Avg_Gas_Consump_m3	Avg_Income_per_inhab_x1000Euro	Prct_Households_low_income	Nr people using wellfare allowance	Nr people unemployment allowance	Nr thefts from house or barn	Nr of cases destructive behaviour	Nr of companies and organisations
0	Amersfoort	155226	2740.0	1130.0	28.1	7.4	3600.0	2660.0	5.0	5.0	14315.0
1	Baarn	24630	2860.0	1570.0	31.1	5.8	400.0	360.0	8.0	4.0	3015.0
2	De Bilt	42846	3130.0	1640.0	33.0	6.3	690.0	610.0	3.0	2.0	4585.0

Data frame 16: Utrecht_Municipality_Data_df

The presence of null-values was checked by applying the info-function and creating a heatmap, but no missing data was present.

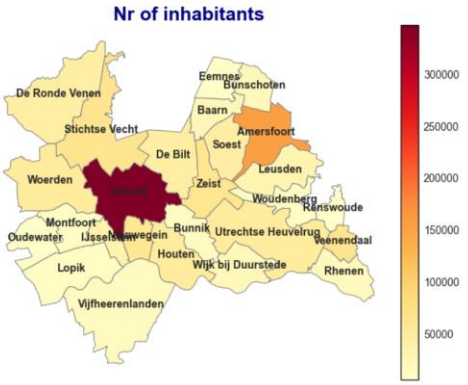
Some additional columns where added where total numbers where turned into relative numbers by dividing the totals by the amount of inhabitants in the municipality and then multiplying the result by 100,000. This was done for the following column variables:

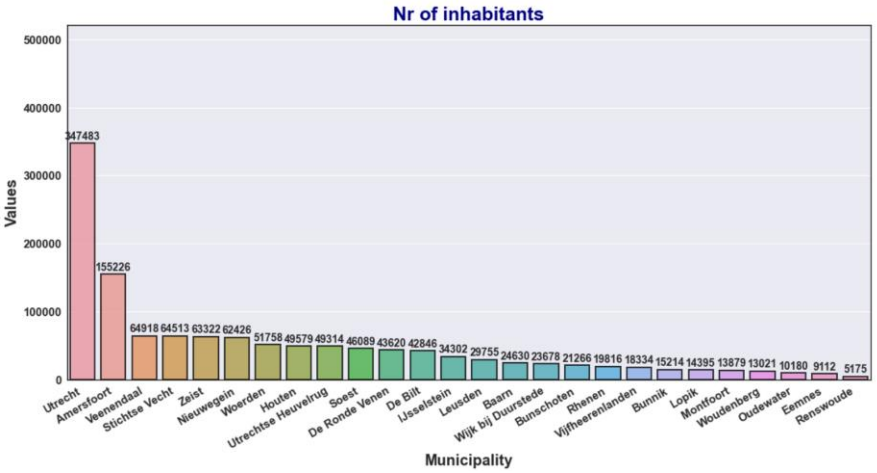
- Nr people using wellfare allowance
- Nr people unemployment allowance
- Nr thefts from house or barn
- Nr of cases destructive behaviour

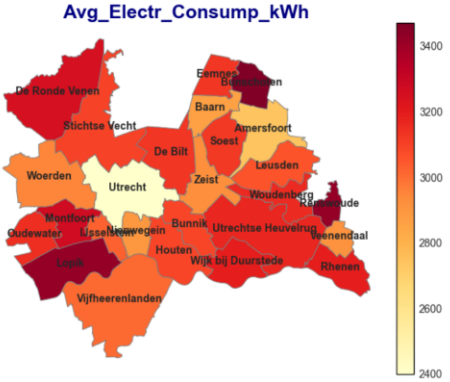
The reason for doing this was because amount of thefts for instance may look like a lot in the province Utrecht when putting it on a choropleth map, however Utrecht also has a lot more inhabitants. Therefore it may not be really good practice to compare absolute numbers of municipalities to each other. Dividing the total number by the amount of inhabitants and multiplying the result by 100,000 may give a more realistic picture and may be a better way to compare one municipality to another.

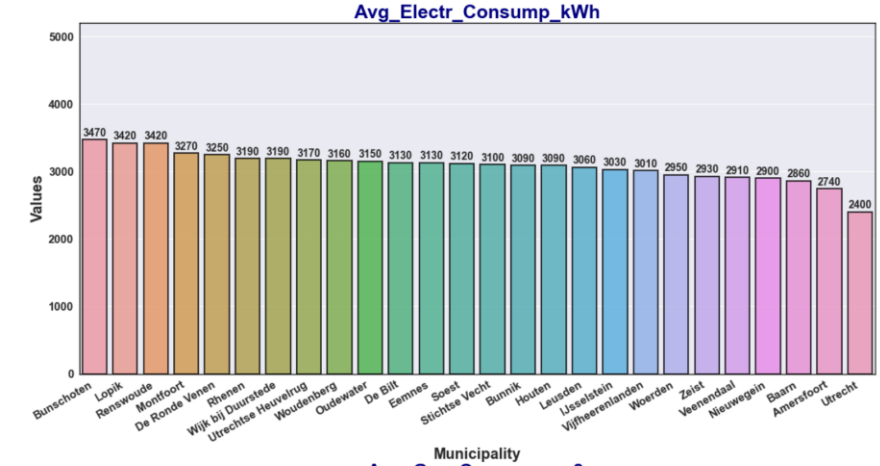
The resulting data frame was then merged to a data frame containing the geospatial data in order to be able to put different figures on a choropleth map.

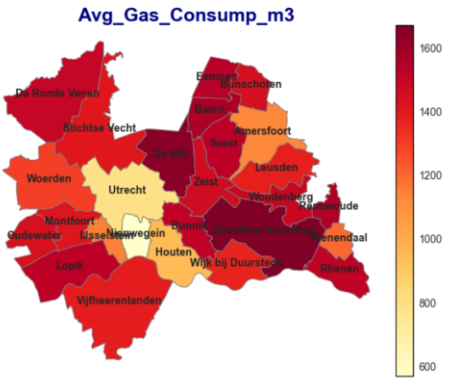
After that, subplots where generated in which both a choropleth map and bar chart were created for each of the column-variables of interest. The results are shown on the next pages.

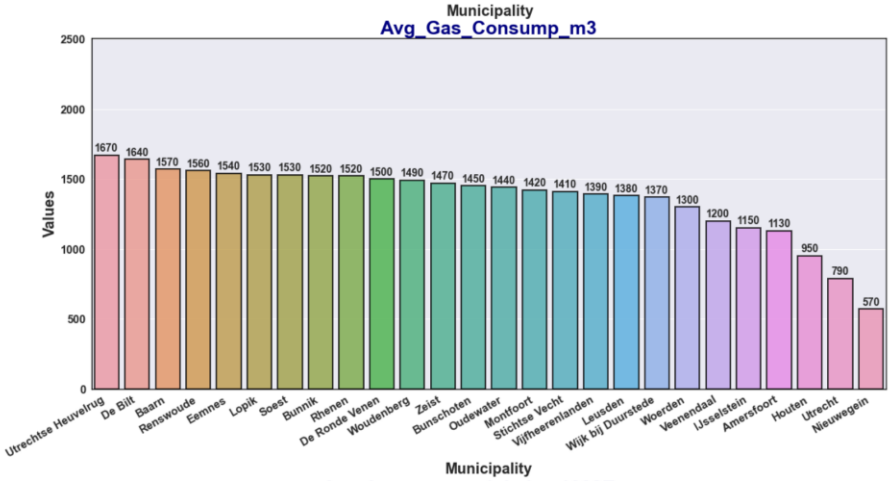

Nr of inhabitants
Eemnes
Bunschoten
De Ronde Venen
Baarn
Stichtse Vecht
Amersfoort
Soest
De Bilt
Leusden
Woerden
Zeist
Utrecht
Woudenberg
Renswoude
Montfoort
Bunnik
Utrechtse Heuvelrug
Oudewater
IJsselstein
Nieuwegein
Veenendaal
Houten
Lopik
Wijk bij Duurstede
Rhenen
Vijfheerenlanden
300000
250000
200000
150000
100000
50000

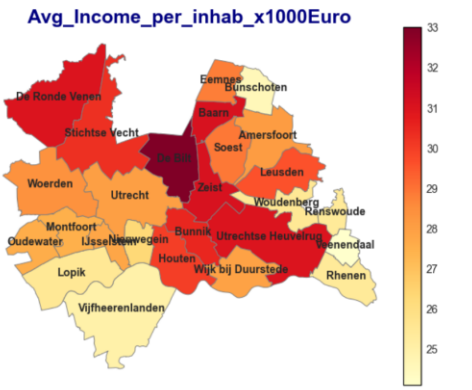

Nr of inhabitants
Values
500000
400000
300000
200000
100000
0
347483
155226
64918
64513
63322
62426
51758
49579
49314
46089
43620
42846
34302
29755
24630
23678
21266
19816
18334
15214
14395
13879
13021
10180
9112
5175
Utrecht
Amersfoort
Veenendaal
Stichtse Vecht
Zeist
Nieuwegein
Woerden
Houten
Utrechtse Heuvelrug
Soest
De Ronde Venen
De Bilt
IJsselstein
Leusden
Baarn
Wijk bij Duurstede
Bunschoten
Rhenen
Vijfheerenlanden
Bunnik
Lopik
Montfoort
Woudenberg
Oudewater
Eemnes
Renswoude
Municipality

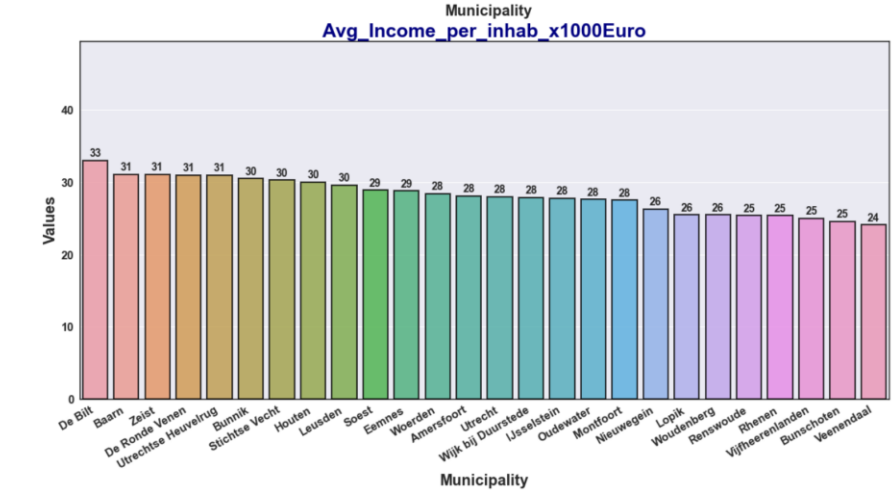

Avg_Electr_Consump_kWh
Eemnes
Bunschoten
De Ronde Venen
Baarn
Stichtse Vecht
Amersfoort
Soest
De Bilt
Leusden
Woerden
Zeist
Utrecht
Woudenberg
Renswoude
Montfoort
Oudewater
IJsselstein
Nieuwegein
Bunnik
Utrechtse Heuvelrug
Veenendaal
Houten
Lopik
Wijk bij Duurstede
Rhenen
Vijfheerenlanden
3400
3200
3000
2800
2600
2400

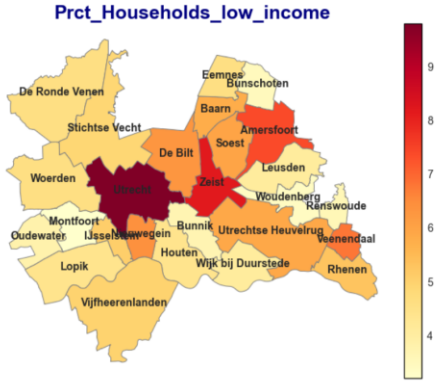

Avg_Electr_Consump_kWh
Values
5000
4000
3000
2000
1000
0
3470
3420
3420
3270
3250
3190
3190
3170
3160
3150
3130
3130
3120
3100
3090
3090
3060
3030
3010
2950
2930
2910
2900
2860
2740
2400
Bunschoten
Lopik
Renswoude
Montfoort
De Ronde Venen
Rhenen
Wijk bij Duurstede
Utrechtse Heuvelrug
Woudenberg
Oudewater
De Bilt
Eemnes
Soest
Stichtse Vecht
Bunnik
Houten
Leusden
IJsselstein
Vijfheerenlanden
Woerden
Zeist
Veenendaal
Nieuwegein
Baarn
Amersfoort
Utrecht
Municipality

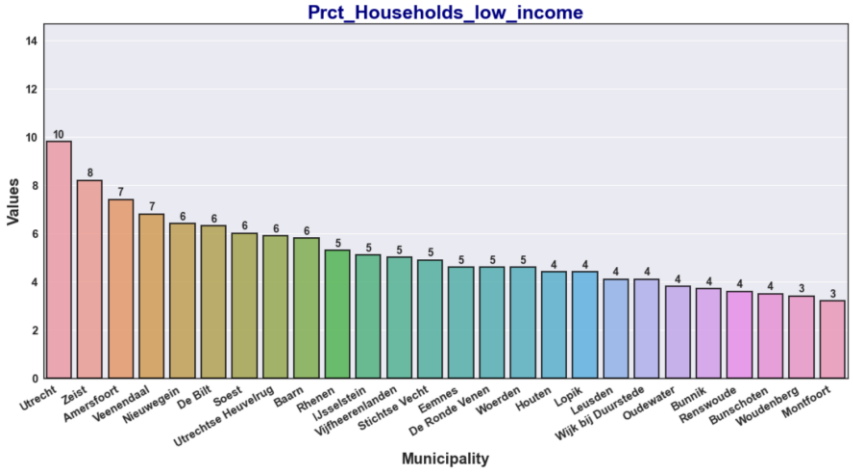

Avg_Gas_Consump_m3
Eemnes
Bunschoten
De Ronde Venen
Baarn
Stichtse Vecht
Amersfoort
Soest
De Bilt
Leusden
Woerden
Zeist
Utrecht
Woudenberg
Renswoude
Montfoort
Oudewater
IJsselstein
Nieuwegein
Bunnik
Utrechtse Heuvelrug
Veenendaal
Houten
Lopik
Wijk bij Duurstede
Rhenen
Vijfheerenlanden
1600
1400
1200
1000
800
600


Avg_Gas_Consump_m3
Values
2500
2000
1500
1000
500
0
1670
1640
1570
1560
1540
1530
1530
1520
1520
1500
1490
1470
1450
1440
1420
1410
1390
1380
1370
1300
1200
1150
1130
950
790
570
Utrechtse Heuvelrug
De Bilt
Baarn
Renswoude
Eemnes
Lopik
Soest
Bunnik
Rhenen
De Ronde Venen
Woudenberg
Zeist
Bunschoten
Oudewater
Montfoort
Stichtse Vecht
Vijfheerenlanden
Leusden
Wijk bij Duurstede
Woerden
Veenendaal
IJsselstein
Amersfoort
Houten
Utrecht
Nieuwegein
Municipality

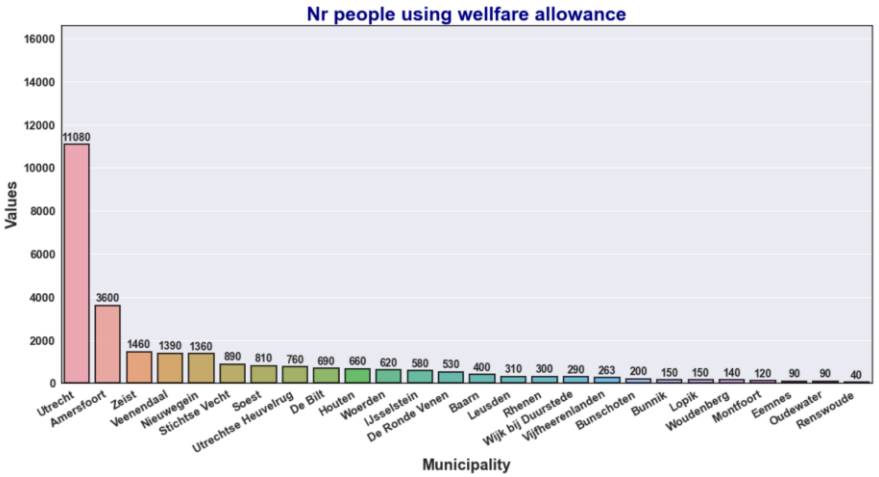

Avg_Income_per_inhab_x1000Euro
Eemnes
Bunschoten
De Ronde Venen
Baarn
Stichtse Vecht
Amersfoort
Soest
De Bilt
Leusden
Woerden
Zeist
Utrecht
Woudenberg
Renswoude
Montfoort
Oudewater
IJsselstein
Nieuwegein
Bunnik
Utrechtse Heuvelrug
Veenendaal
Houten
Lopik
Wijk bij Duurstede
Rhenen
Vijfheerenlanden
33
32
31
30
29
28
27
26
25

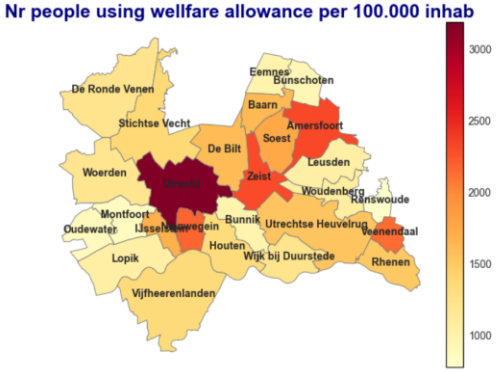

Avg_Income_per_inhab_x1000Euro
Values
40
30
20
10
0
33
31
31
31
31
30
30
30
30
29
29
28
28
28
28
28
28
28
26
26
26
25
25
25
25
24
De Bilt
Baarn
Zeist
De Ronde Venen
Utrechtse Heuvelrug
Bunnik
Stichtse Vecht
Houten
Leusden
Soest
Eemnes
Woerden
Amersfoort
Utrecht
Wijk bij Duurstede
IJsselstein
Oudewater
Montfoort
Nieuwegein
Lopik
Woudenberg
Renswoude
Rhenen
Vijfheerenlanden
Bunschoten
Veenendaal
Municipality

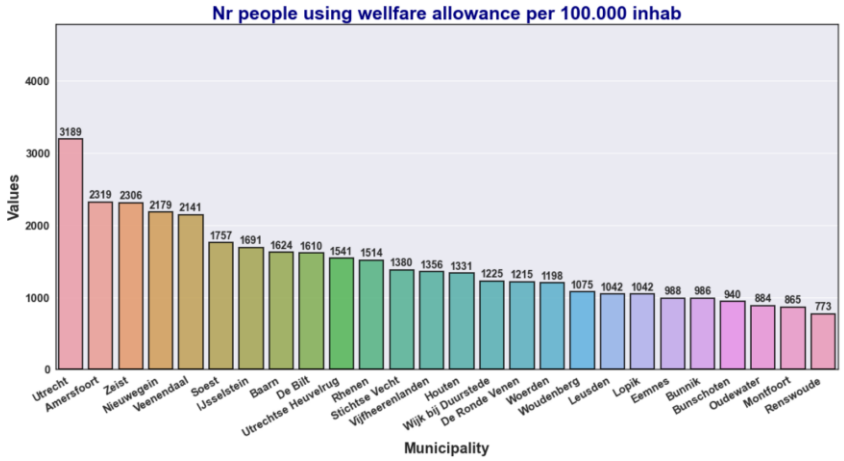

Prct_Households_low_income
De Ronde Venen
Eemnes
Bunschoten
Baarn
Stichtse Vecht
Amersfoort
Soest
De Bilt
Leusden
Woerden
Utrecht
Zeist
Woudenberg
Ranswoude
Montfoort
Bunnik
Utrechtse Heuvelrug
Oudewater
IJsselstein
Veenendaal
Houten
Wijk bij Duurstede
Lopik
Rhenen
Vijfheerenlanden
9
8
7
6
5
4

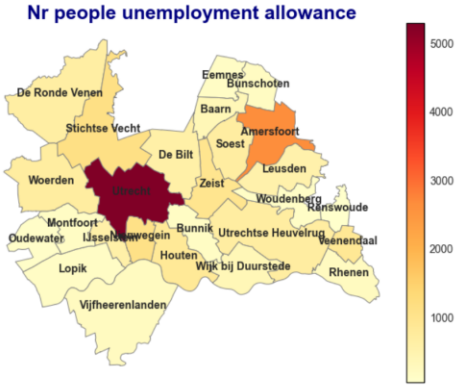

Prct_Households_low_income
Values
Municipality
Utrecht 10
Zeist 8
Amersfoort 7
Veenendaal 7
Nieuwegein 6
De Bilt 6
Soest 6
Utrechtse Heuvelrug 6
Baarn 6
Rhenen 5
IJsselstein 5
Vijfheerenlanden 5
Stichtse Vecht 5
Eemnes 5
De Ronde Venen 5
Woerden 5
Houten 4
Lopik 4
Leusden 4
Wijk bij Duurstede 4
Oudewater 4
Bunnik 4
Renswoude 4
Bunschoten 4
Woudenberg 3
Montfoort 3

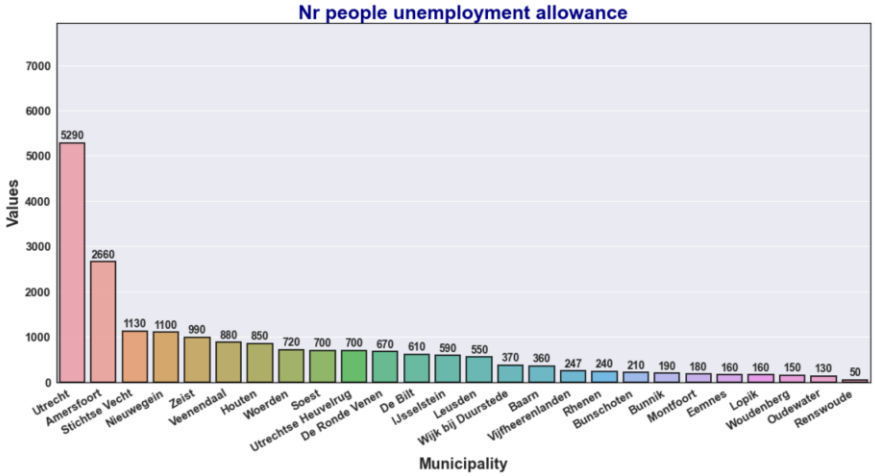

Nr people using wellfare allowance
De Ronde Venen
Eemnes
Bunschoten
Baarn
Stichtse Vecht
Amersfoort
Soest
De Bilt
Leusden
Woerden
Utrecht
Zeist
Woudenberg
Ranswoude
Montfoort
Bunnik
Utrechtse Heuvelrug
Oudewater
IJsselstein
Veenendaal
Houten
Wijk bij Duurstede
Lopik
Rhenen
Vijfheerenlanden
10000
8000
6000
4000
2000


Nr people using wellfare allowance
Values
Municipality
Utrecht 11080
Amersfoort 3600
Zeist 1460
Veenendaal 1390
Nieuwegein 1360
Stichtse Vecht 890
Soest 810
Utrechtse Heuvelrug 760
De Bilt 690
Houten 660
Woerden 620
IJsselstein 580
De Ronde Venen 530
Baarn 400
Leusden 310
Rhenen 300
Wijk bij Duurstede 290
Vijfheerenlanden 263
Bunschoten 200
Bunnik 150
Lopik 150
Woudenberg 140
Montfoort 120
Eemnes 90
Oudewater 90
Renswoude 40

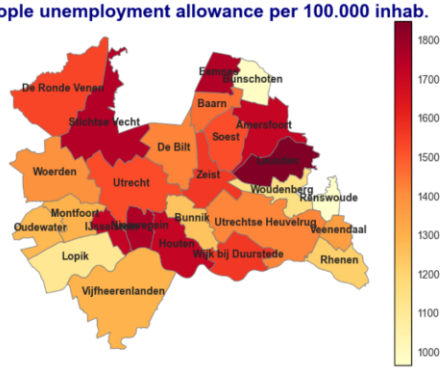

Nr people using wellfare allowance per 100.000 inhab
De Ronde Venen
Eemnes
Bunschoten
Baarn
Stichtse Vecht
Amersfoort
Soest
De Bilt
Leusden
Woerden
Utrecht
Zeist
Woudenberg
Ranswoude
Montfoort
Bunnik
Utrechtse Heuvelrug
Oudewater
IJsselstein
Veenendaal
Houten
Wijk bij Duurstede
Lopik
Rhenen
Vijfheerenlanden
3000
2500
2000
1500
1000

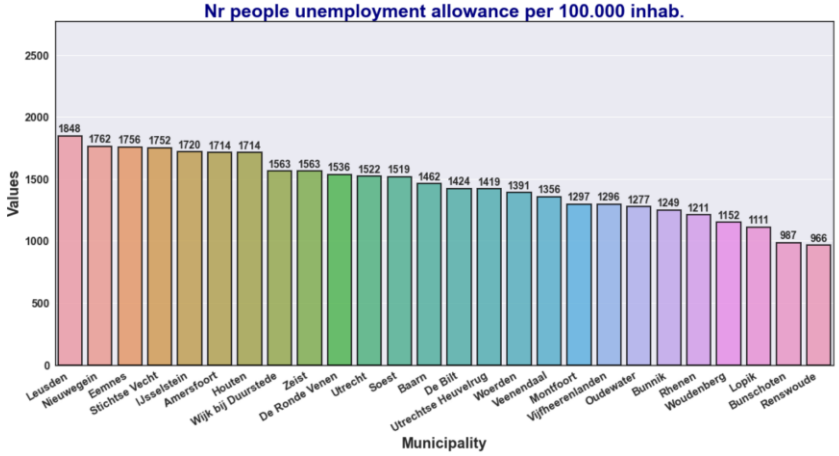

Nr people using wellfare allowance per 100.000 inhab
Values
Municipality
Utrecht 3189
Amersfoort 2319
Zeist 2306
Nieuwegein 2179
Veenendaal 2141
Soest 1757
IJsselstein 1691
Baarn 1624
De Bilt 1610
Utrechtse Heuvelrug 1541
Rhenen 1514
Stichtse Vecht 1380
Vijfheerenlanden 1356
Houten 1331
Wijk bij Duurstede 1225
De Ronde Venen 1215
Woerden 1198
Woudenberg 1075
Leusden 1042
Lopik 1042
Eemnes 988
Bunnik 986
Bunschoten 940
Oudewater 884
Montfoort 865
Renswoude 773


Nr people unemployment allowance
De Ronde Venen
Eemnes
Bunschoten
Baarn
Stichtse Vecht
Amersfoort
Soest
De Bilt
Leusden
Woerden
Utrecht
Zeist
Woudenberg
Ranswoude
Montfoort
Bunnik
Utrechtse Heuvelrug
Oudewater
IJsselstein
Veenendaal
Houten
Wijk bij Duurstede
Lopik
Rhenen
Vijfheerenlanden
5000
4000
3000
2000
1000

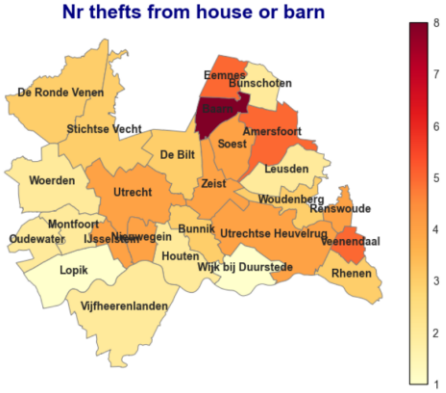

Nr people unemployment allowance
Values
Municipality
Utrecht 5290
Amersfoort 2660
Stichtse Vecht 1130
Nieuwegein 1100
Zeist 990
Veenendaal 880
Houten 850
Woerden 720
Soest 700
Utrechtse Heuvelrug 700
De Ronde Venen 670
De Bilt 610
IJsselstein 590
Leusden 550
Wijk bij Duurstede 370
Baarn 360
Vijfheerenlanden 247
Rhenen 240
Bunschoten 210
Bunnik 190
Montfoort 180
Eemnes 160
Lopik 160
Woudenberg 150
Oudewater 130
Renswoude 50

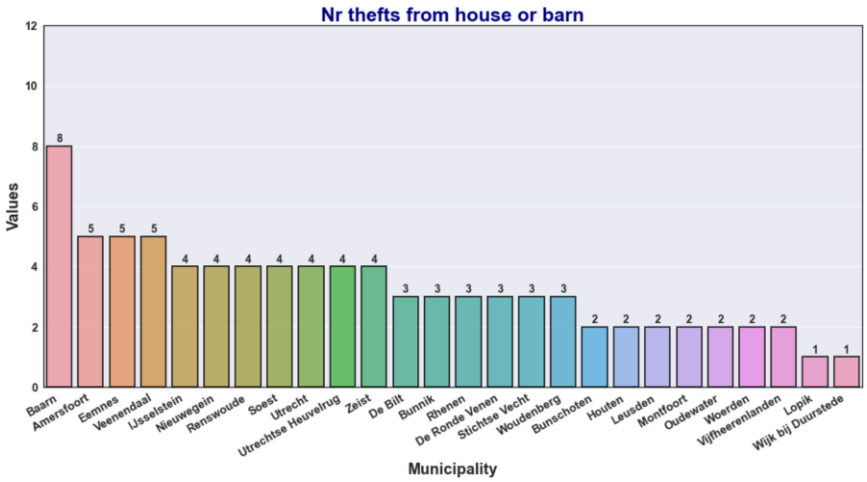
Nr people unemployment allowance per 100.000 inhab.



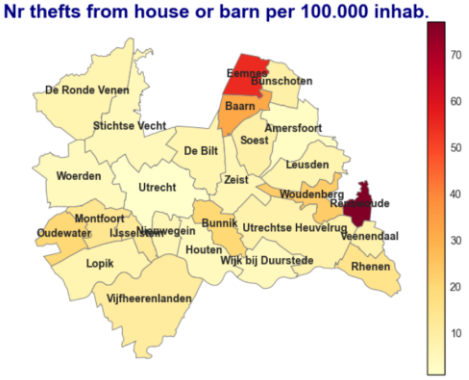


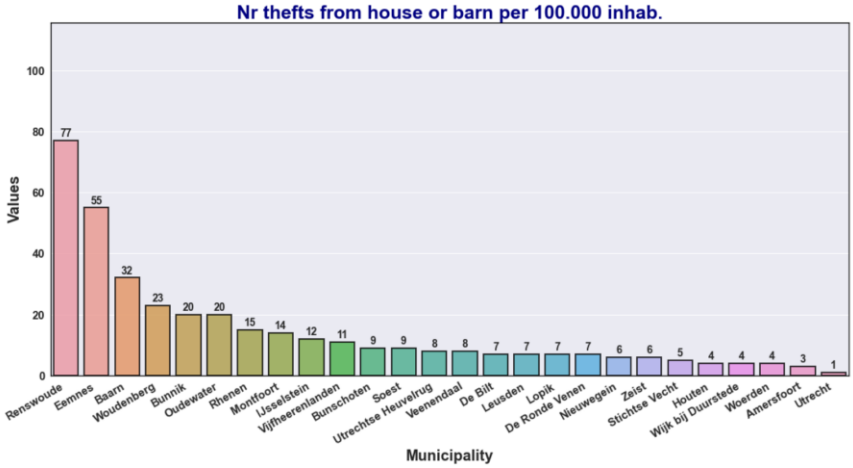
Nr thefts from house or barn



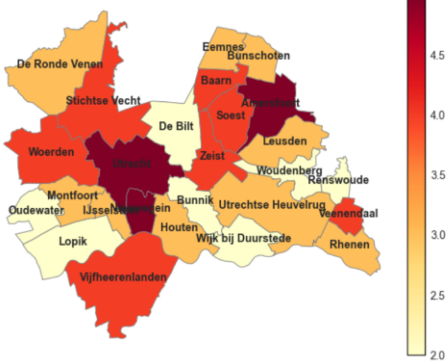


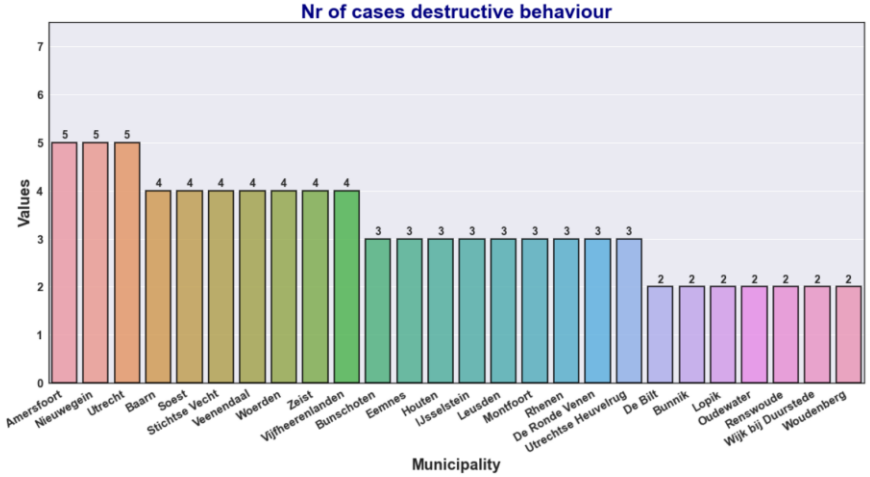
Nr thefts from house or barn per 100.000 inhab.

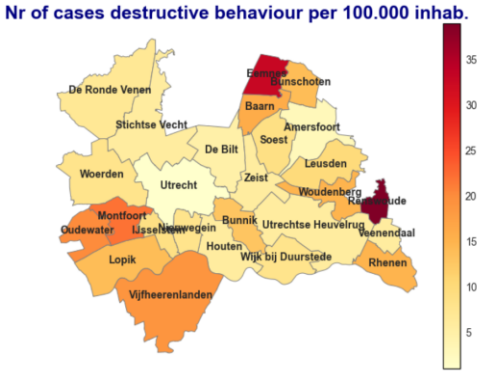


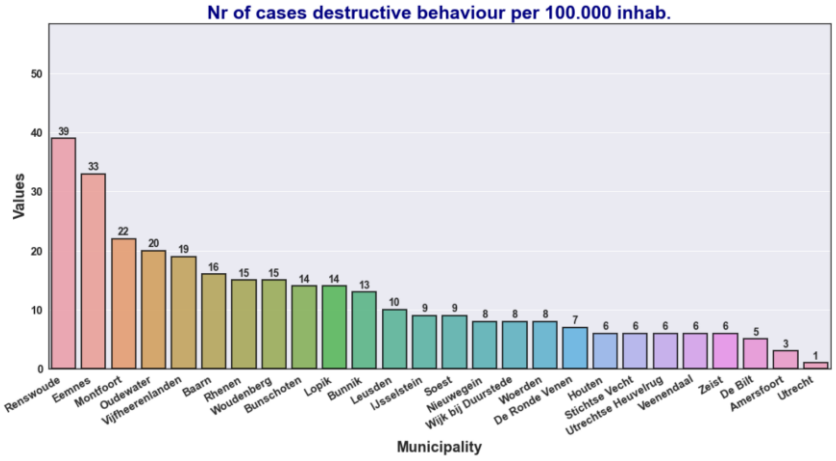


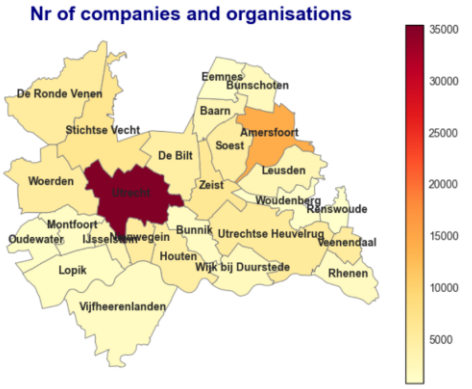
Nr of cases destructive behaviour

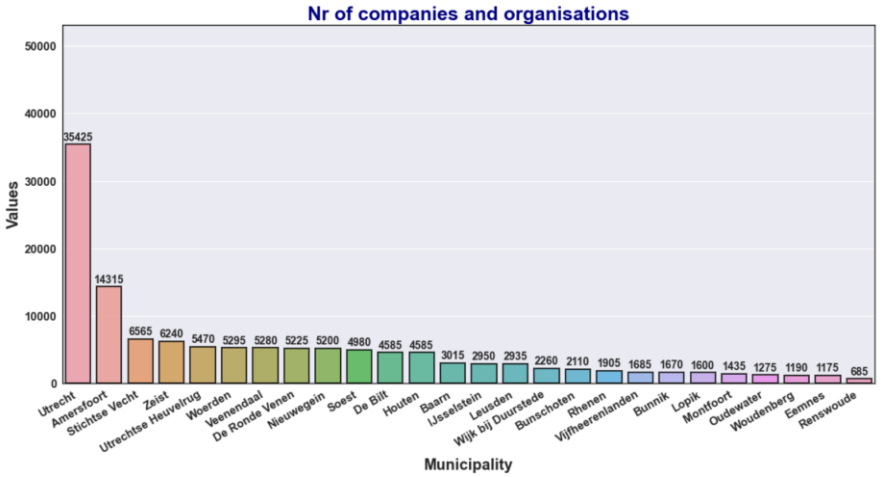












Observations and conclusions about the above choropleth maps and bar charts is included in the next chapter in which the results will be discussed.

5. Results

In context of the business problem, potential interesting areas to invest in real estate had to be found in the municipality of Utrecht. A couple of criteria/assumptions were given for 'potential interesting areas' by the fictional real estate agency, which are:

- House prices over the past 5 years have a stable increasing trend in the area of interest;
- Ideally the house price is currently reasonably low, but has a historical trend that indicates that house prices in that area will rise further in the future.
- There is a variety of venues nearby, such as supermarkets, restaurants, shopping malls, etc.
- Some additional factors that can be beneficial:
  - Low crime rates / nuisance figures;
  - Low energy consumption;
  - Low number of people using employment allowance;
  - High number of organisations nearby (employment opportunities)

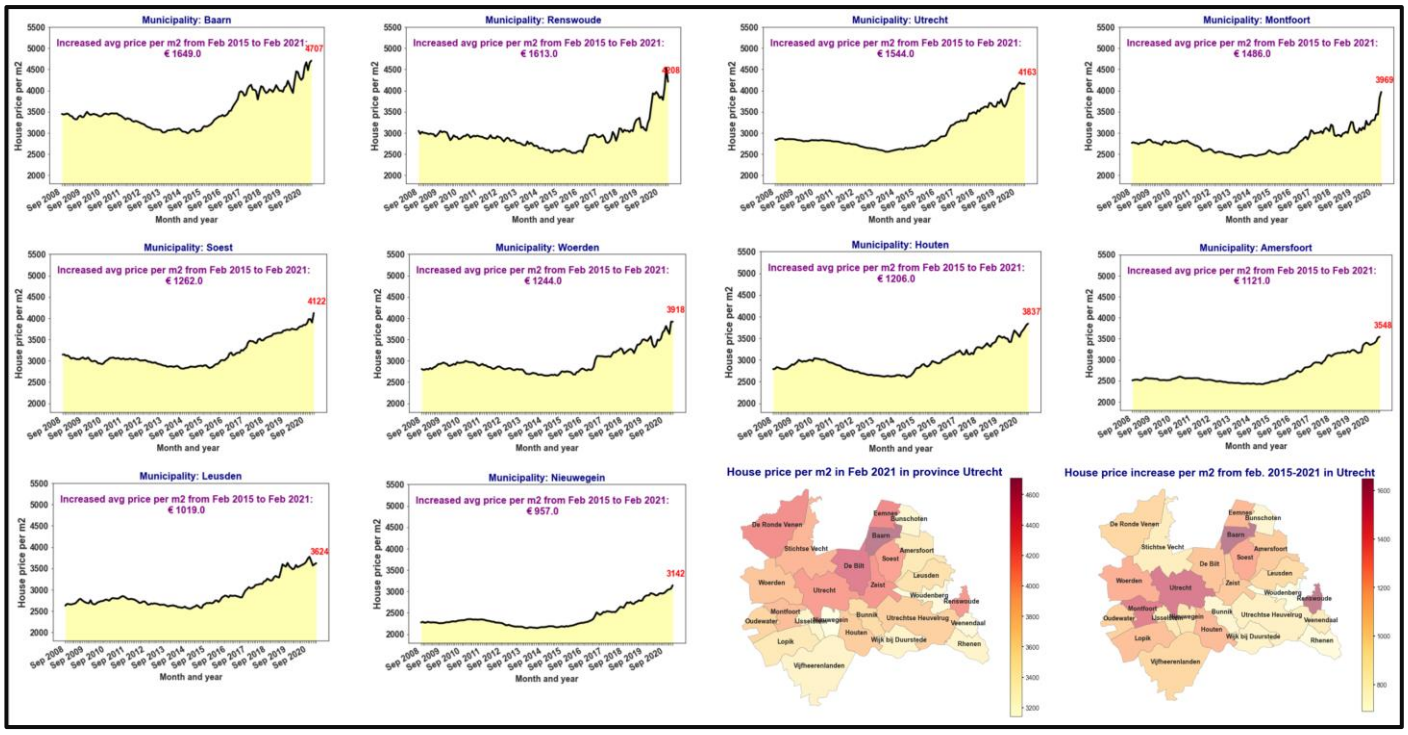
Putting all the information together with above criteria in mind, the following can be concluded from the data and visualisations that were made.

<u>REQUIREMENTS 1 AND 2</u>:

1. **House prices over the past 5 years have a stable increasing trend in the area of interest**
2. **Ideally the house price is currently reasonably low, but has a historical trend that indicates that house prices in that area will rise further in the future.**

To get answers for requirements 1 and 2, line charts and choropleth maps where created, showing the trends for house prices per m^2 for each of the 26 municipalities in Utrecht. The charts also show the most recent house price value per m^2 and has a note on top of the chart with the price increase over the past 6 years. Based on these graphs the following top 10 municipalities seemed to be potentially interesting areas for investing in real estate:

- Baarn
- Renswoude
- Utrecht
- Montfoort
- Soest
- Woerden
- Houten
- Amersfoort
- Leusden
- Nieuwegein



When looking at the choropleth maps, potential interesting choices could be municipalities where house prices are relatively low and increase in house prices have been high. For the above 2 choropleth maps, this means a particular municipality has a more yellowish colour on the left choropleth map and a more orange/reddish colour on the right choropleth map. The following municipalities would then be interesting areas:

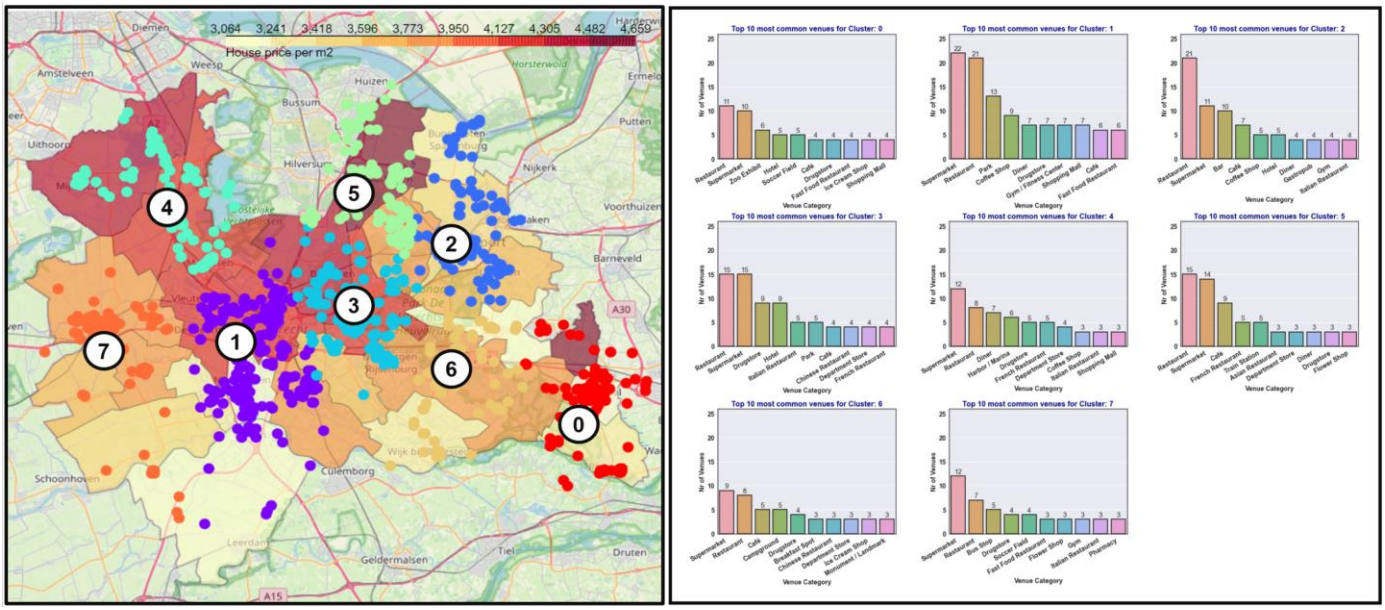
- Montfoort
- Ijsselstein
- Nieuwegein
- Amersfoort
- Woerden

The above mentioned municipalities seem to have low- to average house prices and have had a stable increase in house price per m^2 over the past few years.

REQUIREMENT 3:

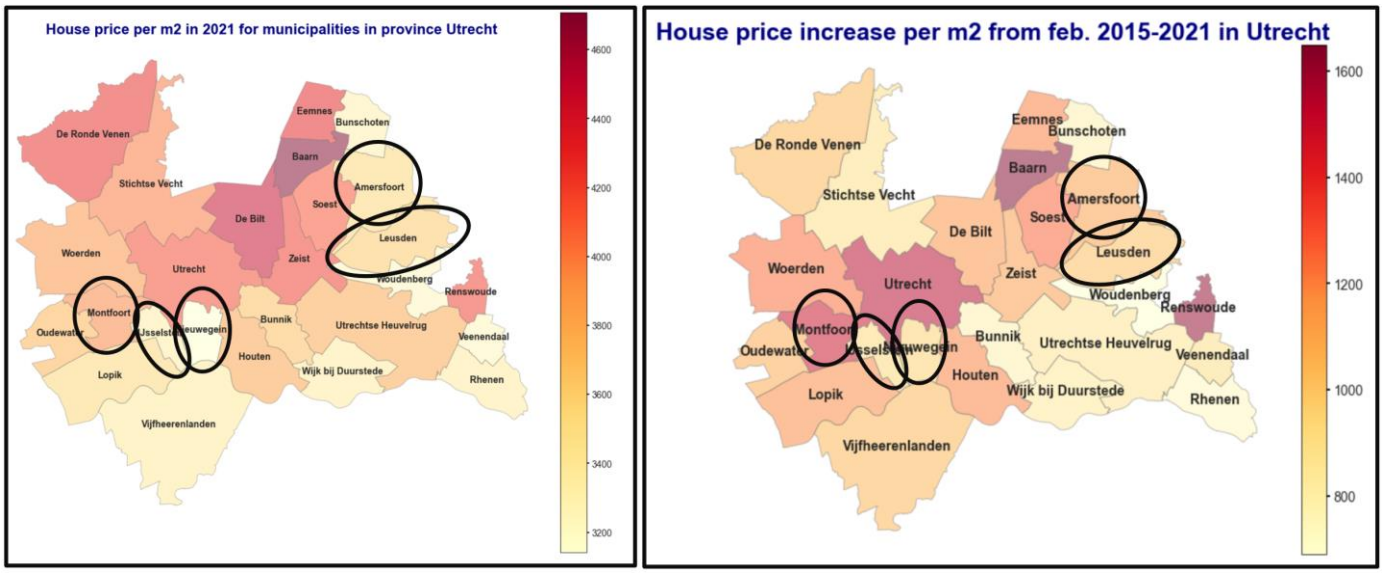
3. There is a variety of venues nearby, such as supermarkets, restaurants, shopping malls, etc.

By using the Foursquare API we obtained location data from venues within a 5km radius from the centre point of each municipality. The K-means algorithm was applied on the coordinates of the venues and clusters were formed based on how near the venues were to each other. The most suitable amount of clusters 'k' was found to be 8, which was established using the 'elbow' method that plots a distortion score against the amount of clusters 'k'. After clusters were created by the K-means algorithm, the array of clusters was included in the data frame with venue names. Then a choropleth map was made using the 'folium library'. The map shows the average house prices per m^2 per municipality. The venues were plotted on top of the choropleth map by markers (dots). Clusters were visualized by varying colour of the markers. Bar charts were created for each cluster, showing the top 10 most common venues in the cluster in descending order. The result is shown below.



Overall there seems to be multiple clusters and sub areas in those cluster where it could be interesting to invest in real-estate with the aim of reselling it for a higher price. Cluster 1, 3, 4 and 5 are in the highest house price range and have a high degree of urbanity, with dense areas of venues being closely packed together. Around those clusters there are some areas that also have a high venue count and density, but where house prices are lower, which is the 'south part of cluster 1' and 'cluster 2'. Assuming that we aim for areas with low house prices and potential high profit in the future, then the south part of cluster 1 and cluster 2 would be interesting areas to examine further.

The interesting areas are circled in black below:



The current situation in the Netherlands (March 2021) is that there is a great shortage of house offerings on the market. House prices have risen so much that young people cannot get a mortgage to purchase their first house. Waiting for a period until house prices reduce, which can take a while, is often not a choice that young couples that want to settle and start a family are willing to take. An increasing number of people therefore look for places a little further away from the main city, where house prices are a little lower. There is currently also a trend ongoing of older people who move away from the city centres to purchase a larger free-standing-house in more remote, cheaper areas to enjoy nature and the peaceful landscape outside the city. As demand for accommodations outside the city increase, house prices outside the city will also increase. Considering this, the above encircled areas that are a little outside the main city of Utrecht are therefore likely to yield good profit when investing and reselling houses in those areas.

ADDITIONAL FACTORS THAT CAN BE INTERESTING WHEN SELECTING POTENTIAL AREAS FOR INVESTING IN PROPERTY:

- **Low crime rates / nuisance figures;**
- **Low energie consumption;**
- **Low number of people using unemployment allowance;**
- **High number of organisations nearby (employment opportunities)**

Data related to above 4 mentioned themes were obtained via API and the library 'cbsodata' from the Dutch Bureau of Statistics. A collection of subplots was generated, where choropleth maps and complementary bar charts where shown. These charts can be found at the end of previous chapter. Some interesting findings/insights gained from the choropleth maps and bar charts are:

- The municipality Utrecht has by far the most inhabitants (347,483 in the chart) and has more than twice the amount of inhabitants compared to the second largest city Amersfoort (155,226 inhabitants). The third municipality with most inhabitants is Veenendaal having 64,918 inhabitants, which is significantly lower compared to Utrecht and Amersfoort.
- Both the average electricity and average gas consumption in the municipalities Utrecht and Amersfoort are less compared to the other municipalities in the province. I found an article on the internet that claimed that the cause of the lower energy consumption is that the main cities have a lot more apartments than free-standing-houses. Heat loss and living areas in the apartments are less compared to free-standing houses and hence the larger amount of apartments in main cities result in a lower average for energy consumption. Another factor for lower energy consumption in main cities is the presence of district heating. Households that use district heating commonly do not consume any gas.
- Average income seems to be higher in some higher-priced municipalities near the cities Utrecht and Amersfoort and municipalities between Utrecht and Amsterdam. The top 10 municipalities with highest

average income are: De Bilt, Baarn, Zeist, De Ronde Venen, Utrechtse Heuvelrug, Bunnik, Stichtse Vecht, Houten, Leusden and Soest. When looking at the choropleth maps, we see that house prices for most of the mentioned municipalities are in the higher spectrum (orange to reddish colour). Exceptions are the municipalities Leusden and Utrechtse Heuvelrug, where house prices are in the lower- to average spectrum (yellow- to light orange on the map), but where average income seems to be higher.

- Contrary to what might be expected, the percentage of households with low income seems to be higher in cities with higher average house prices. The top 10 cities with the highest percentage of households with a low income as indicated by the CBS are: Utrecht, Zeist, Amersfoort, Veenendaal, Nieuwegein, De Bilt, Soest, Utrechtse Heuvelrug, Baarn and Rhenen.
- The relative number of people using financial governmental support (welfare allowance, unemployment allowance) seem to be high for municipalities in the vicinity of main cities, such as Utrecht and Amersfoort. Also in areas around Utrecht there seem to be areas where the relative number of people using financial support is high. There is not really a clear area where there are multiple municipalities that have a lower relative number of people using unemployment allowances. When looking at absolute numbers, then Utrecht and Amersfoort have most people using unemployment allowances, but these municipalities have a lot more inhabitants than the other municipalities and therefore the comparison to other municipalities using absolute numbers would not be meaningful.
- For the relative number of thefts and registered cases of destructive behaviour of people per municipality, some areas seem to be more popular, which are: Renswoude, Eemnes, Baarn, Woudenberg, Bunnik and Oudewater. In Renswoude, Eemnes and Baarn the house prices per m^2 are relatively high. Apparently some people in the theft-business seem to be aware of peoples wealth in those areas. However – looking at the total number of thefts we see that the amount is relatively small.
- The top 10 number of municipalities with highest number of companies/organisation are: Utrecht, Amersfoort, Stichtse Vecht, Zeist, Utrechtse Heuvelrug, Woerden, Veenendaal, De Ronde Venen, Nieuwegein, Soest. Areas close to potential jobs tend to be more in demand than places that are more remote and require more travel time to the work location.

Having looked at above additional information, the following can be said:

- Looking at the number of people living in each municipality, we can see that Utrecht and Amersfoort are the most popular areas to live.
- The fact that average incomes are higher in and near the cities is meaningful to know, but is not very surprising as living in cities is more expensive and there are more jobs available. The fact that in Leusden average house price is relatively low, but average income is relatively high is interesting. People that can spend more money on their home could drive up the house prices in that area.
- When looking at the absolute (total) numbers for theft and registered cases for destructive behaviour by people, the numbers are very low compared to the number of inhabitants in each municipality.
- Unsurprisingly most registered companies/organisations are found in the municipalities Utrecht and Amersfoort, which are also the two main cities in province Utrecht. Living in or near those areas can be beneficial in terms of employment opportunities, which attract potential house buyers.

6. Discussion

In the previous section, the findings from the choropleth maps and bar charts in the methodology section were discussed.

In this section we will discuss the observations that were made and recommendations that can be made. The assignment was to search for potential interesting areas within the province Utrecht to invest in real estate on the residential house market. A couple of preconditions were given by the fictional real estate investment organisation to refine the search for interesting areas, which were:

- House prices over the past 5 years have a stable increasing trend in the area of interest;
- Ideally the house price is currently reasonably low, but has a historical trend that indicates that house prices in that area will rise further in the coming future.
- There is a variety of venues nearby, such as supermarkets, restaurants, shopping malls, etc.

Having taken the above criteria into consideration, the following municipalities were selected as potentially interesting areas to purchase and resell properties:

- Amersfoort
- Leusden
- Nieuwegein
- Ijsselstein
- Woerden

The above selected municipalities have low- to average house prices, venues are positioned relatively close together and venue count and variety is relatively high, also house prices in these areas have had a stable increasing trend over the past few years and these municipalities are in or near the main cities Utrecht and Amersfoort, where house prices are high.

I should note that my area of expertise is not in real estate and in reality the best choice may simply be to purchase and resell properties in the areas where demand for accommodations is highest and that the magnitude of the house price does not play a role. Therefore I deliberately specified some preconditions at the beginning of this assignment that seemed logical criteria when looking for interesting areas to invest in property. From some news articles that I read, I knew that in the current situation of the Dutch house market, there is a great shortage of houses, which is driving up the house prices further. A lot of people that want to purchase their first house are not eligible for a mortgage because their income does not match the needed loan they would need to purchase the house. As a result there is an increasing trend of people looking for cheaper houses outside the cities, but still relatively close to the city areas. There is also a trend going on of older people that choose to leave the city area to purchase larger free-standing houses in more remote areas. The corona crisis may speed up these trends as well. As more people have started working from home during the corona crisis, companies were forced to ensure that the required technologies that allow people to work at home were in place. During and after the crisis this may change the way we work permanently and the need to be near a city centre may become less as well, because people can work remotely. As a result, demand for houses outside the city may rise, which in turn would drive up the house prices outside the city.

By taking these developments into consideration it seems to be a good choice to look for places near the cities where house prices are still relatively low, where there is a variety of venues near each other and where there has been a stable upward trend in house prices. The 5 selected municipalities met the specified criteria. Therefore my recommendation to the real estate agency would be to focus on those municipalities to purchase and resell property with the aim of making a profit.

A challenge I faced when making this report was related to applying the K-means algorithm. Initially I had put different information together in a data frame, including variables with information about neighbourhoods, such as crime rates, average income, distance to large supermarkets, distances to restaurants, distances to cafes, etc. After normalizing the data and applying the k-means algorithm clusters were formed. However - after studying the clusters I found that almost all the data points were appointed to one particular cluster and only a few to 2 clusters. A lot of the variables had outliers in the

data and that could have affected the outcome of the k-means algorithm. After working on the dataset for a while and looking for manners to yield a more plausible separation into clusters, I chose to only use the coordinates of the venues to apply the k-means clustering algorithm. The elbow method helped to select the most suitable number for the amount of clusters 'k'. However when plotting the data on the choropleth map using the folium library, I thought that the dataset could be split up in more clusters. Therefore I also applied the Silhouette Coefficient method and drew a Silhouette Diagram to examine if another value for the amount of clusters may be better. I found that the highest Coefficient was at k = 41, so 41 clusters. However - when applying the Coefficient Method the number for the ideal amount of clusters changed in the Silhouette Method. Therefore I was not convinced that k = 41 was a reliable number for choosing the ideal amount of clusters. Eventually I chose the amount of clusters to be 8, because that was the amount of clusters that the elbow method indicated to be the ideal amount of clusters and this also corresponded to the first peak in the line-chart for the Silhouette Coefficient.

I obviously still have a lot to learn about programming in Python and about algorithms that are used in data science, when those are useful and how to feed those models with the right (pre-processed) data that can be turned into useful information. However - the journey of discovering the fundamentals of data science, learning the basics of programming in Python and coming up with a self-chosen challenge was interesting. Especially during the last assignment I have learned a lot, as I searched the internet to apply certain techniques/syntax in Python in order to process data or turn data into a certain table or graphic. Hopefully - the things I learned in this course will form a good basis for further developing my knowledge and skills in analysing data and applying data science techniques.

7. Conclusions

The objective was to find specific areas in Utrecht that would be interesting to invest in houses with the aim of celling and reselling those to make a profit. Data from different sources and in different forms were gathered from the internet using different python libraries, such as Pandas, Geopandas, Foursquare API and the API from the Dutch Central Bureau of Statistics using the library 'cbsodata'. Data was examined, pre-processed and merged together to form data frames that could be used to make choropleth maps with geopandas and folium and to create line charts and bar plots that gave us the required information to draw conclusions about potential interesting investment areas, thereby taking into account a few preconditions set by the real estate investment agency. Eventually, 5 of the 26 municipalities met the criteria, which were **Amersfoort, Leusden, Nieuwegein, Ijsselstein, Woerden**. For these municipalities, house price trends have had a relatively stable increasing trend over the past few years. Besides that currently the average house prices are relatively low in these municipalities. There is also a variety of venues at or near these municipalities and all municipalities are in or near larger cities, which seem to become more popular places, especially to people that purchase their first house.