Frederic Bärtl & Nicolas Gaillard
Zürich & Bern
frederic.baertl@students.unibe.ch
nicolas.gaillard@students.unibe.ch

Data Science Project

Correlation of rent prices with two variables in Switzerland - Conceptual Design Report

31st October 2022

Abstract

Rental prices are a recurring topic in Swiss newspapers. We hear about steady increases in the Region of Zurich but no explanations are provided why such a trend is isolated to certain areas or why certain cantons are less prone to rental prices increases. We therefore though about the the most obvious factors that might influence such a trend and came up with the research question:

Are rent prices in Switzerland correlated with the ratio of empty housing and the population density? The testing of this hypothesis has been performed by using data aggregated on a cantonal level by the Bundesamt für Statistik. While we manage to give a few first answers, this project led us to formulating a couple more questions than we had at the start. Indeed, even if the stated correlation may seem pretty obvious, one would need to have more disaggregated data as well as information on many more variables to be able to test this hypothesis in a proper scientific way.

In the end, this project was mainly an opportunity for us to test the techniques we had learned during the first two modules of the CAS ADS 2022. This was also a very good first experience of working with real data and tackling the different challenges linked to it.

Table of Contents

1 Project Objectives

This project and our research question follows the capitalistic assumption of supply and demand in real estate and its application in Switzerland. As the supply and demand mechanisms are very complex and could not be looked at during this project, we simplified their definitions: the percentage of empty flats is the supply and the population density is the demand. We are well aware that flats which are not occupied are not necessarily available for occupation and that other factors can influence the supply (insalubrious flats, bad location, etc.). We are also well aware that population density does not necessarily mean that there is a higher demand, as families for example live more densely but are not looking for more flats.

Thus, through this project, our goal is simply to determine if the evolution of the rent prices is correlated with the percentage of empty flats and the population density percentage per Swiss canton. As this correlation seems to be pretty obvious at first view, the purpose of this project is mainly to give us the opportunity to apply the knowledge we acquired during the first two modules of the CAS ADS 2022.

We expect to be able to determine the strength of the aforementioned correlation but also to confront ourselves with the challenges of handling real data.
For this, we need to use histograms, scatter plots as well as QQ plots.

2 Methods

We intend to use sets of real data available through the website of the Bundesamt für Statistik under the format of the result of surveys conducted by this institution. This means that the data has already been preprocessed and that there is a high probability for the available data to only be values resulting from the analysis of the collected data. We thus expect to work mainly with means and time series (mean rent price for 2022 in canton Bern for example).

The advantage of using this kind of data in our expectation is that it has already been collected, cleaned and prepared for a statistical analysis. There should be a reduced need of filtering out redundant and noisy data, of unifying the data formats and of transforming the data.

As we work in a team, we will be working on the Jupyter environment of Google Colab. We will import the different datasets into Pandas DataFrames and will be using the following Python libraries and modules:

- `pandas`, which is the most common open-source Python library made for working with relational data in various data structures;

- `io`, which provides Python's main facilities for dealing with various types of inputs and outputs;
- `numpy`, which offers comprehensive mathematical functions;
- `matplotlib.pyplot`, which provides an implicit way of plotting and is intended for simple cases of programmatic plot generation;
- `scipy.stats`, which is a module of the `scipy` library containing a huge number of statistical functions;
- `traitlets`, which is a library allowing classes of objects to have different attributes;
- `statsmodels.formula.api`, which is an interface for specifying models using formula strings and DataFrames.

We will be using standard deviation to evaluate the data spread as well as regression to evaluate the relationship between our variables. As our dataset will be fairly small (one measure per variable per 27 cantons), we will not have to determine a sample size.

We also plan to use some normality tests on our most important variables using the D'Agostino-Pearson method. We might also perform an analysis of variance to determine if our findings are statistically significant. However, as we will be looking at different time series, we are yet unsure about the correct hypothesis testing we should perform and will thus look at this problem later during our project.

3 Data

We extracted different datasets from bfs.admin.ch and combined the date for the different datasets and for the different time frames into a single dataframe (see CSV below). For this we used information from the following sources:

- Average rental costs per square meter split by rooms (per appartement) and split by canton from 2012 to 2020 [1]
- Empty quota of apartments ("Leerstand" in German) per canton from 2012 to 2020 [2]
- Density of the people living in a apartment per room as well as per appartement [3]

By taking a first look at those three datasets, we see that the data is in .xslx format, is split per time frame in single sheets and has noise (additional information about the data per time frame).

As writing a script to solve the three issues we just mentioned would be pretty complex due to factors like: noise per sheet, changes in formatting over time, changes in reporting, difference in highlighting missing/incomplete information. Consequently we will perform the data preparation manually. We first merge the different time frames into one, filter out the noise before converting the file to csv [4].

This is an example of of the rows of the final file look like:

	Region	Year	Total	1Room_rent_SQM	2Room_rent_SQM	3Room_rent_SQM	4Room_rent_SQM	5Room_rent_SQM	6Room_rent_SQM
0	Schweiz	2020	16.5	20.4	18.2	16.6	15.6	15.6	16.5
27	Schweiz	2019	16.4	19.3	18.0	16.5	15.6	15.4	16.1
54	Schweiz	2018	16.2	19.2	17.7	16.3	15.5	15.3	16.3
81	Schweiz	2017	15.9	18.2	17.2	16.0	15.2	15.3	16.0
108	Schweiz	2016	15.8	18.1	17.2	15.9	15.1	15.1	16.1
135	Schweiz	2015	15.6	17.6	16.8	15.7	15.1	15.0	16.3
162	Schweiz	2014	16.2	18.0	17.2	16.3	15.7	15.6	16.6
189	Schweiz	2013	16.0	17.6	17.2	16.1	15.5	15.5	16.7
216	Schweiz	2012	15.9	18.3	17.3	16.0	15.4	15.1	15.9

Table 1: Example of the Overview of the different arrays present in the data

4 Metadata

As we use relational data, the metadata required for reproducing our analysis is as follows:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 243 entries, 0 to 242
Data columns (total 20 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Region                 243 non-null    object
 1   Year                   243 non-null    int64
 2   Total                  243 non-null    float64
 3   1Room_rent_SQM         243 non-null    float64
 4   2Room_rent_SQM         243 non-null    float64
 5   3Room_rent_SQM         243 non-null    float64
 6   4Room_rent_SQM         243 non-null    float64
 7   5Room_rent_SQM         243 non-null    float64
 8   6Room_rent_SQM         243 non-null    float64
 9   Empty_quota            243 non-null    float64
 10  Density_p_per_room     243 non-null    float64
 11  density_p_per_flat     243 non-null    object
 12  Size_per_p_in_sqm      243 non-null    object
 13  Ownership_quota        243 non-null    object
 14  Amount_total           243 non-null    int64
 15  amount_Rentals         243 non-null    int64
 16  amount_Genossenschaft  243 non-null    int64
 17  amount_own_flat        243 non-null    int64
 18  amount_own_house       243 non-null    int64
 19  amount_Others          243 non-null    int64
dtypes: float64(9), int64(7), object(4)
memory usage: 38.1+ KB
```

This metadata is stored in the dataframes which are themselves built on the merged dataset that we upload at the beginning and which is accessible via GitHub as well as via our Notebook.

5 Data Quality

We expect the quality of our data to be particularly high as it is being provided by the Bundesamt für Statistik. Indeed, the data should be accurate, complete, reliable and consistent.

Nevertheless, as the data is not accessible in a raw format and as it has been necessary to adapt it, we expect to have at least a minimum of quality loss, which we take into account.

Furthermore, we understand that the data is an aggregation per canton and the amount of underlying data differs between cantons and types of apartments. Therefore the BFS indicated that certain information is less reliable than others (due to smaller sample sizes). A fact that we will not be able to account for in our analysis but need to keep in mind when interpreting the results of specific regions.

6 Data Flow

As we are using data provided by the Bundesamt für Statistik, our data source is not the original one but the results of the studies they have performed. Below you can find a simple data flow (Figure 1).

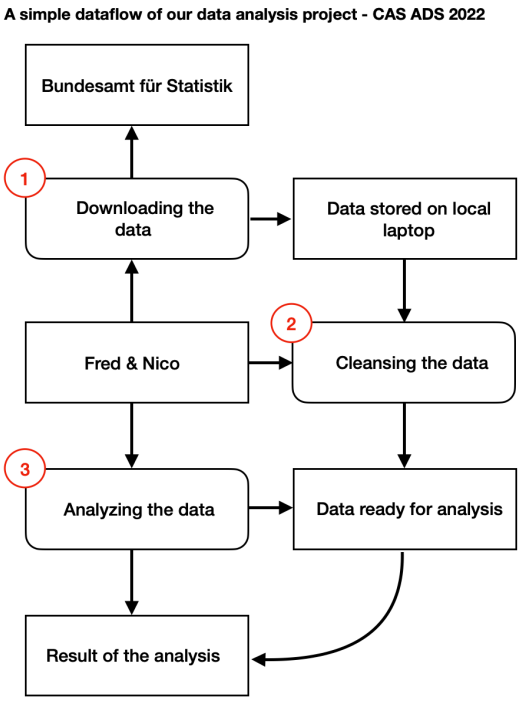

Figure 1: A simple dataflow of our data analysis project

After having downloaded the available relevant data from the website of the Bundesamt für Statistik [1][2][3] and stored it on our laptops, we cleansed it.

The cleansing included the replacement of missing observations (initially indicated with “x”) with 0. We understand that this might imply a biais and the more accurate replacement formula should be taking the average of the observation of the year before and the year after to get a more realistic value. As we intend to use those “assumed” observations, we kept them at “0” to be able to filter them out more easily.

After that, we analyzed the data in order to present the result of our analysis.

7 Data Model

Below one can find the conceptual, logical and physical data models of our project.

Conceptual data model

In the conceptual data model, we sketch out the variables of our data set as well as their relationship. This gives us a chance to gain an overview of our dataset without getting concerned with its details in the first place.

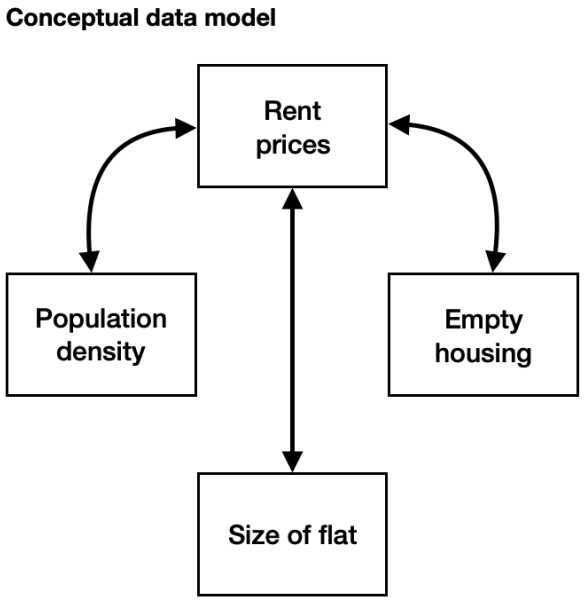

Figure 2: Conceptual data model

Logical data model

In the logical data model, we are supposed to sketch out the data object, their attributes and the relationships between them.

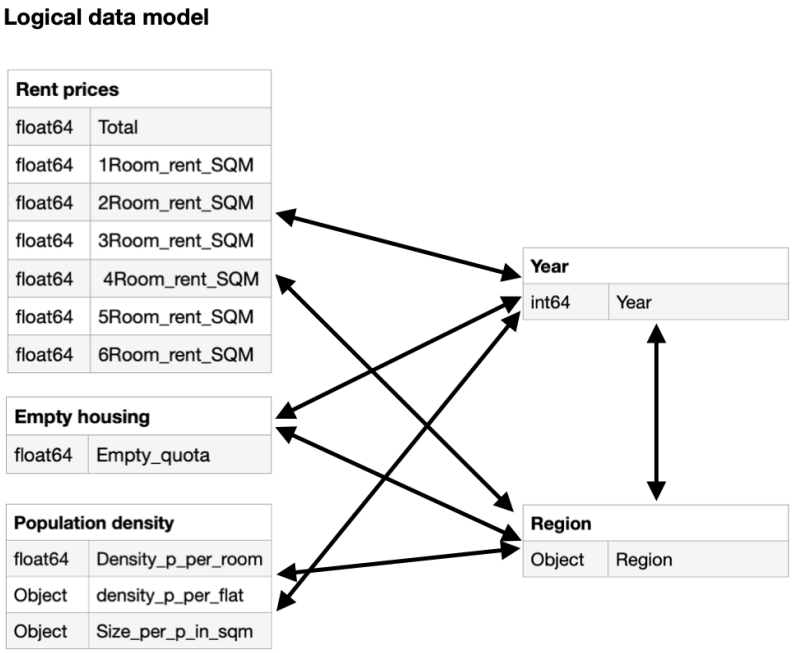

Figure 3: Logical data model

Physical data model

As the physical data model is the final stage of modeling before building the database, we first need to answer the question of the size of that database. Given the fact that the database of this project is going to be very small (values for 27 cantons over eight years), we do not think that a physical data model is appropriate for this project.

8 Risks

The major risk here is that the data is too much aggregated to be able to test the hypothesis in a satisfactory way. If this is the case, the project time schedule would be massively delayed, as we would first need to get access to more detailed data.

We would need to request disaggregated data from the BFS and thereby run into additional problems as the information on rental prices would no longer be on the same level as the information on density and empty houses. Even if all information would be available on the same disaggregated level, a local comparison would be much more complex (e.g. which municipalities

to aggregate back into one “miniregion). Consequently, this would also mean an increase of the project time required as well as the project cost. THe project would at least be put on hold for the time of additional data gathering.

On the other hand, this also means that the risk of a quality drop of our aimed output is very low. Indeed, we are going to work with the results of a data analysis which has been performed by a state agency and officially published.

Nevertheless, another existing risk is the fact that the preparation of the data that we will perform might lead to new errors in the expected output. (e.g. incomplete information on specific observations or false information that has been “absorbed” by the aggregation) In this case, we will have to perform a thorough check of the data to eliminate those errors. This might have a minor impact on the project time schedule.

9 Preliminary Studies

A first preliminary step in the analysis consists of finding out which are the five cantons with the highest rent prices for the last year for which data is available, 2020. For this we simply sorted the cantons per rent price for 2020 (see Figure 4 below).

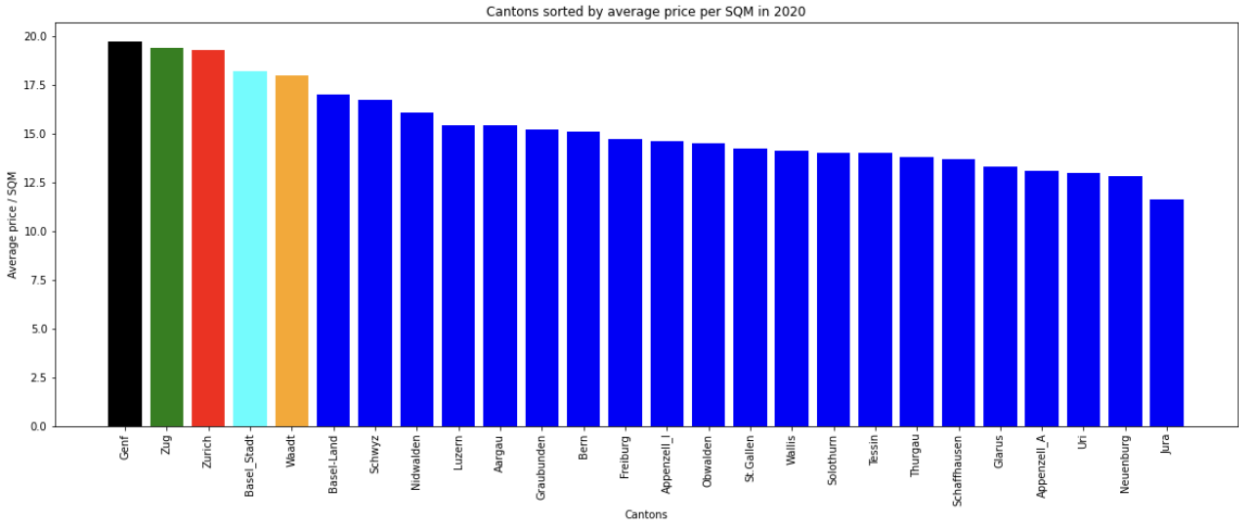

Figure 4: Cantons sorted by average price per square meter in 2020

After this preliminary step we performed another one to see if those five cantons stayed the most expensive throughout the entire available time frame from 2012 to 2020, which can be seen in the figures 5, 6, 7 and 8 below.

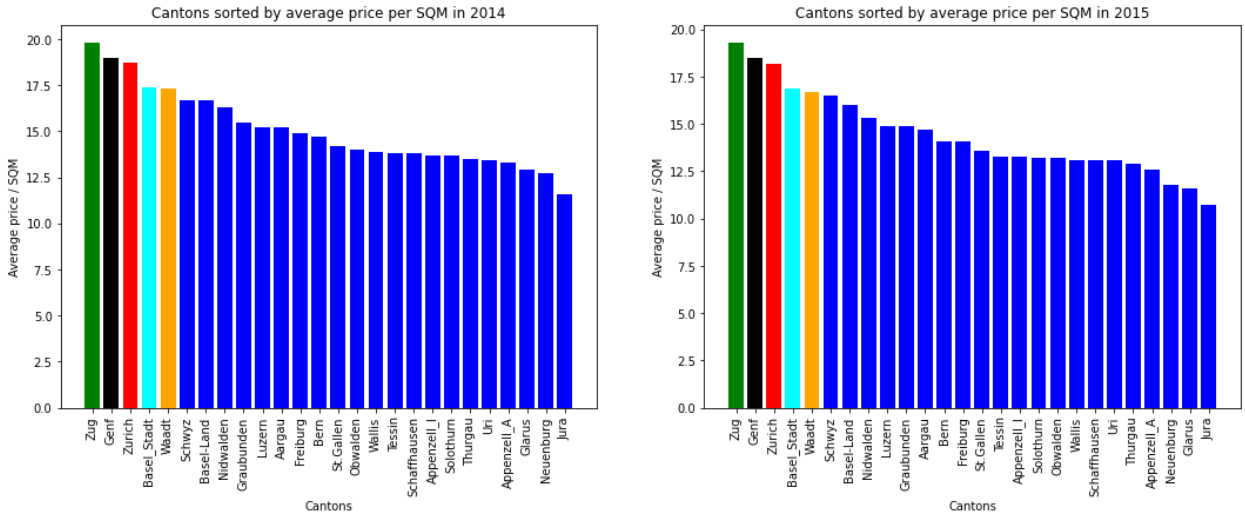

Figure 5: Cantons sorted by average price per square meter in 2012 and 2013

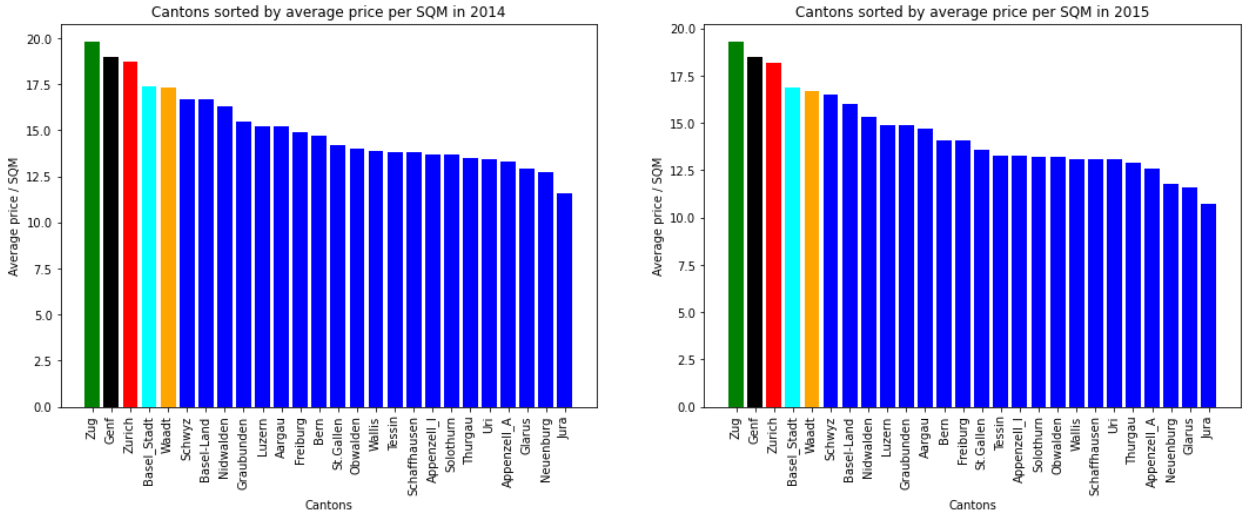

Figure 6: Cantons sorted by average price per square meter in 2014 and 2015

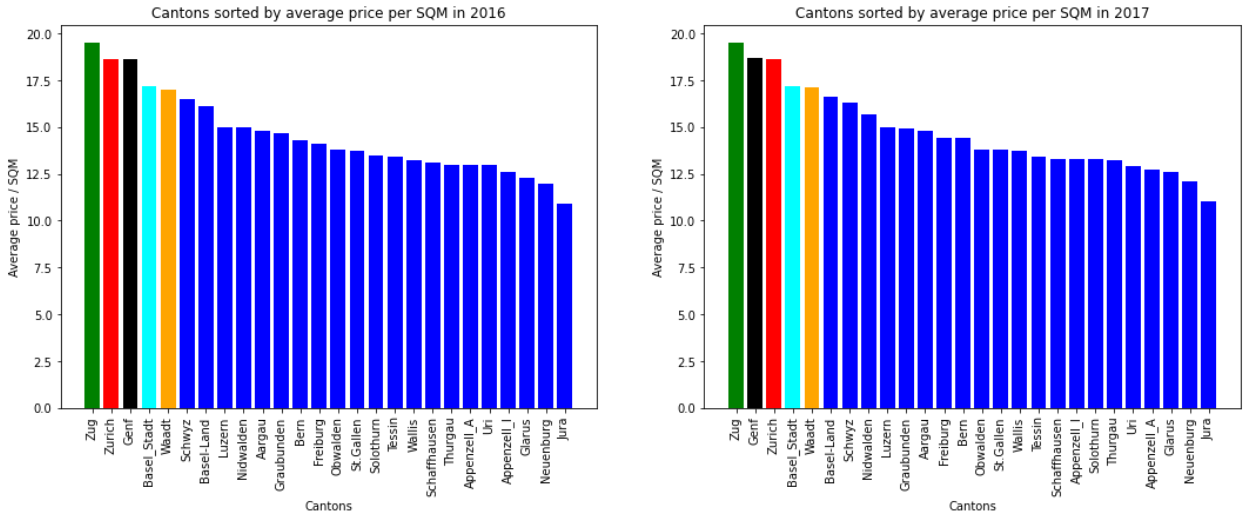

Figure 7: Cantons sorted by average price per square meter in 2016 and 2017

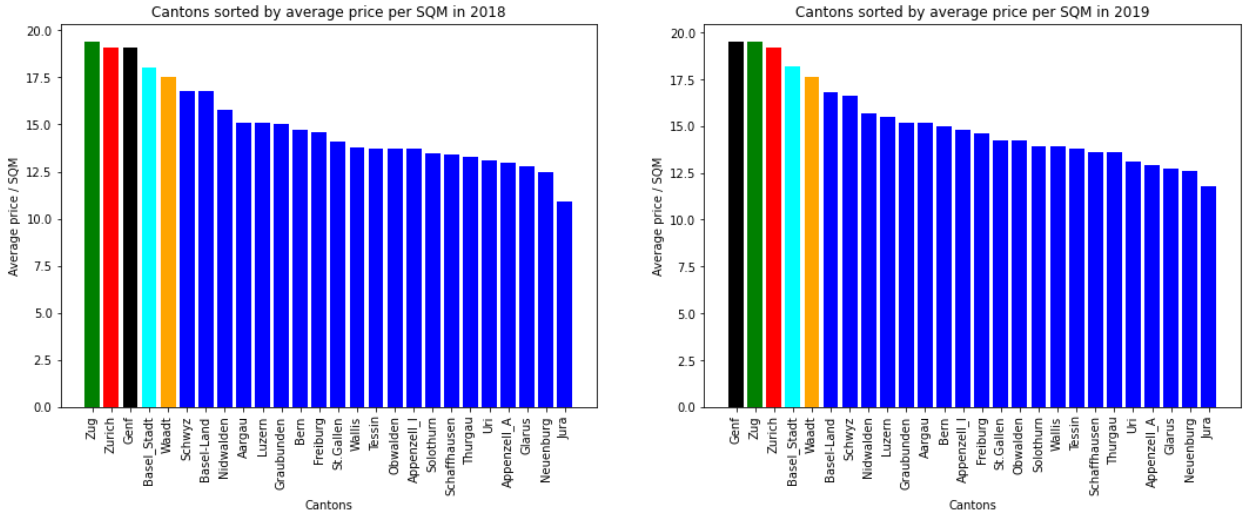

Figure 8: Cantons sorted by average price per square meter in 2018 and 2019

As one can see in the figures above, there has been little modification of the ranking of the cantons per average square meter price. The only variation which is of interest for us is the downgrade of the canton Waadt to the sixth place in 2012.

The next step of our preliminary analysis has been to identify the development of the rent prices for these five top cantons (Genf, Zug, Zurich, Basel Stadt and Waadt) from 2012 to 2020. As shown in figure 9 below, there is a slight increase for Waadt, Basel, Zurich and Geneva while Zug looks more like a stagnation with a slight decrease over time. While the average price per square meter in Switzerland (all 27 cantons) is lower, we see a similar trend as for the 4 cantons (excl. Zug) with an increase from 2012 to 2014 followed by a short decline and a second steady increase from 2015 to 2020.

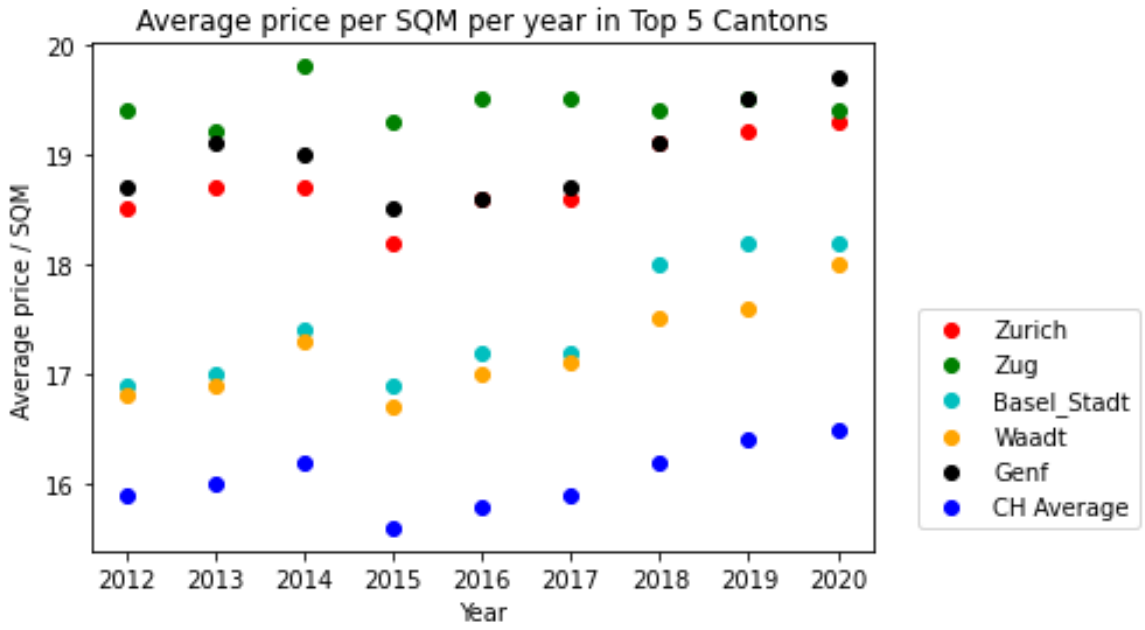

Figure 9: Average rent price per square meter from 2012 to 2020 for the Top 5 cantons and Switzerland

Until this point of the preliminary analysis, we used the "total" rent price ignoring the size of apartments and using average values. This might be misleading considering the difference between 1 room apartments and 6+ room apartments.

As we believe that rooms are a good indication for different segments: 1 room apartments tend to be more basic compared to 6+ room apartments that tend to be more luxurious.

We will now take a look at the evolution of rent prices distinguished by the amount of rooms in the Top 5 cantons by comparing the average price per square meter of one room apartments as well as of two rooms apartments with the average price per square meter in Switzerland (see Figures 10 and 11 below).

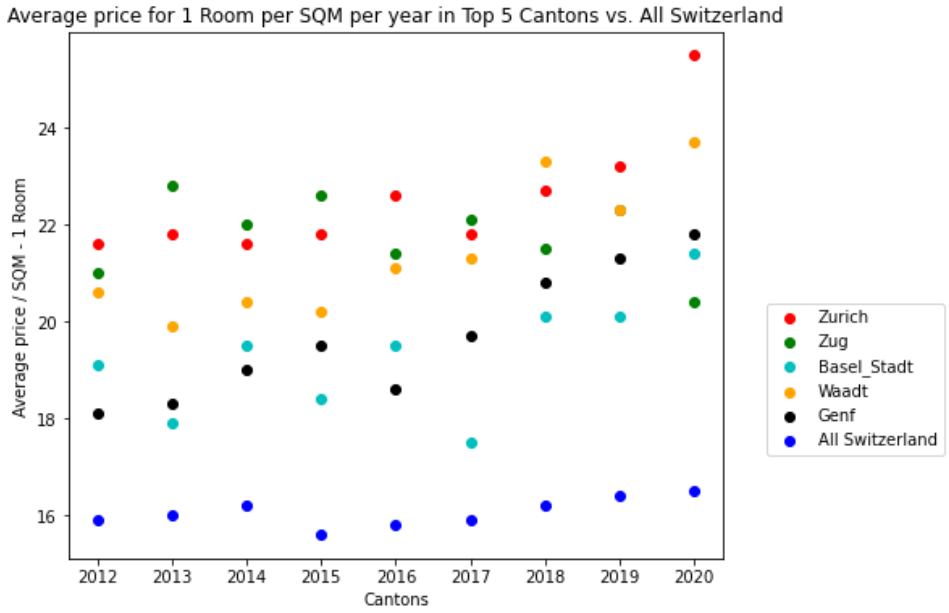

Figure 10: Average rent price per square meter for one room apartments from 2012 to 2020 for the Top 5 cantons and Switzerland

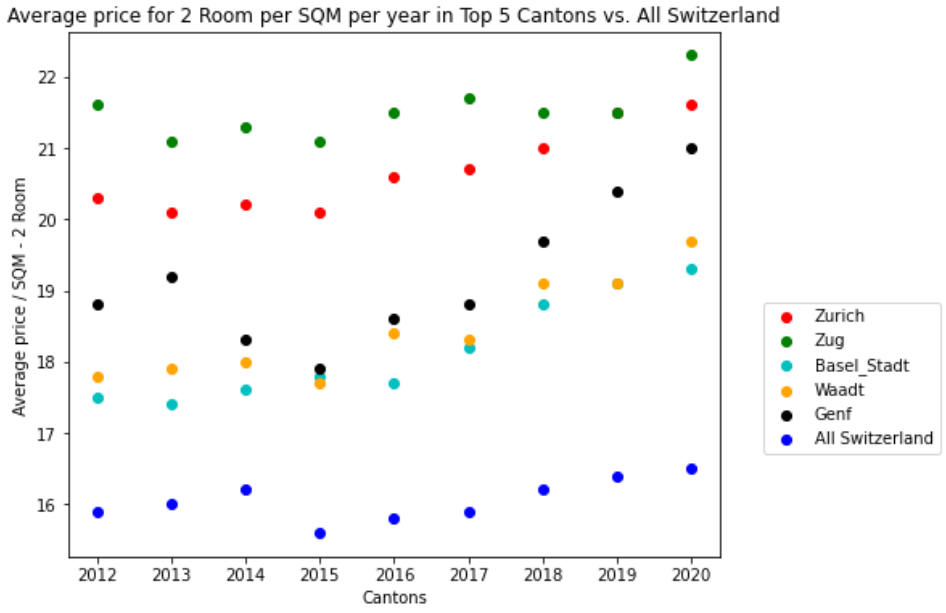

Figure 11: Average rent price per square meter for two rooms apartments from 2012 to 2020 for the Top 5 cantons and Switzerland

As we can see in the figures above, the rent price increase in one room apartments is actually very strong indicating that smaller apartments are more strongly affected by price changes than the "average" (Figure 9). Not only do we see a strong increasing trend for our top 5 cantons

(compared to the small increase of Switzerland as a whole), we also see very high absolute rental prices per sqm for 1 room apartments.

Considering the fact that one or two room apartments are designed for singles in a specific age category, one could try to explain this by a change in the Swiss lifestyle. Nevertheless, this would need to be analyzed in more detail using demographic indicators and is not subject of this analysis.

We now want to see if the rent price trend for the five and six rooms and above apartments is the same as the one seen above for smaller apartments.

We can see above that the trend for 5 Room appartements is less strong but still increasing for certain regions (Figure 12). Fluctuations are much stronger and certain regions tend to see decreasing trends in the last 2-3 years.

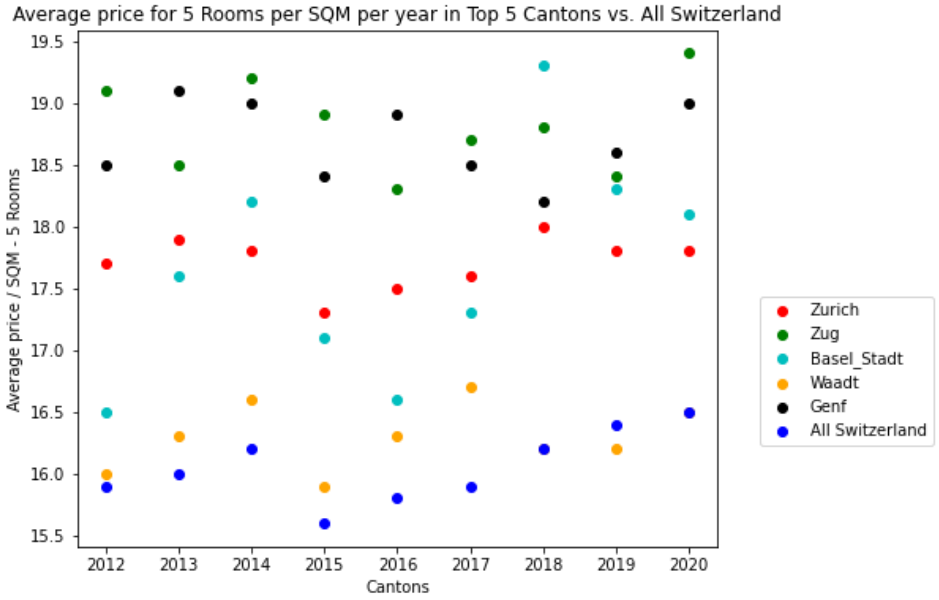

Figure 12: Average rent price per square meter for five rooms apartments from 2012 to 2020 for the Top 5 cantons and Switzerland

By comparison, as shown in the figure below (Figure 13), we see that the trend for "luxury" apartments (six rooms and above) is much weaker with slight increases in Zurich but relatively strong decreases in rental price in Geneva, the most expensive canton. It is unclear if this is due to the fact that availability of this kind of flats in the market increased as we did not incorporate supply information (such as construction specific to six room+ apartments) to our data.

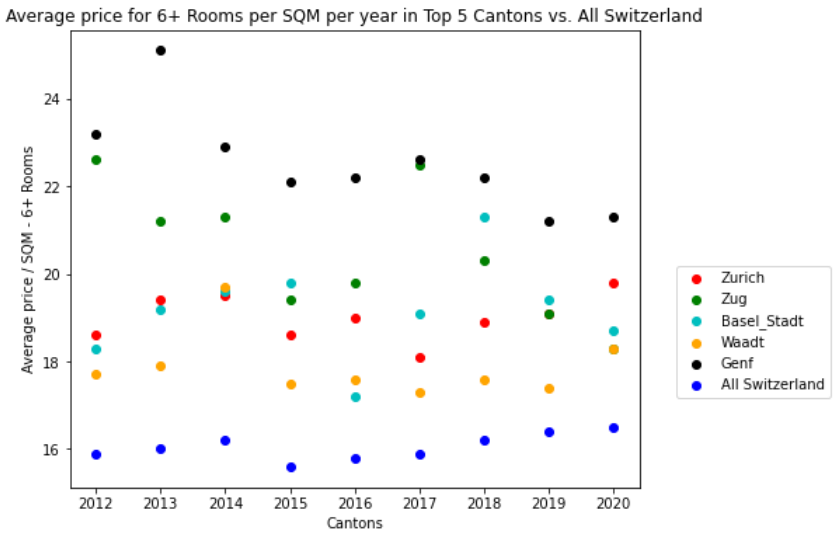

Figure 13: Average rent price per square meter for six rooms and above apartments from 2012 to 2020 for the Top 5 cantons and Switzerland

This preliminary analysis shows that looking at the supply of flats with more rooms alone is not enough to explain higher rent prices. The observations are actually contradicting our assumption that luxury apartments would see stronger price increases. Nevertheless, the reason might not be supply and demand but the false assumption that we derive the level of luxury from the amount of rooms (e.g. a 1 room penthouse might be very luxurious).

A better way to distinguish apartments might be to look at their size as well as at the density of people living in the apartment (per room).

Another assumption might be that high prices are explained by high demand, thus we should expect smaller appartment size (per person) and higher density of people for the top 5 cantons with the highest price levels.

For this, we compare the apartment size in 2020 of the five most expensive regions with the five least densely populated cantons, namely:

- Appelzell-A
- Appenzell-I
- Jura
- Glarus
- Thurgau

Region	Year	Size_per_p_in_sqm
Appenzell_I	2020	50.7
Appenzell_A	2020	51.6
Jura	2020	48.2
Glarus	2020	50.3
Thurgau	2020	51.9
Genf	2020	36.9
Zug	2020	47.6
Zurich	2020	44.9
Basel_Stadt	2020	41.5
Waadt	2020	43.3

Table 2: Average living space in square meters in a flat per person for the Top 5 and Bottom 5 Swiss Cantons for the year 2020

We can see in the table above (Table 2) that, for 2020, while an average person in Geneva occupied 36.9. square meters, a person in Appenzell occupied more than 50 square meters.

If we now compare the density per room of the Top 5 and Bottom 5 cantons, we can see that it is actually substantially higher for the most expensive cantons (see figure 14 below).

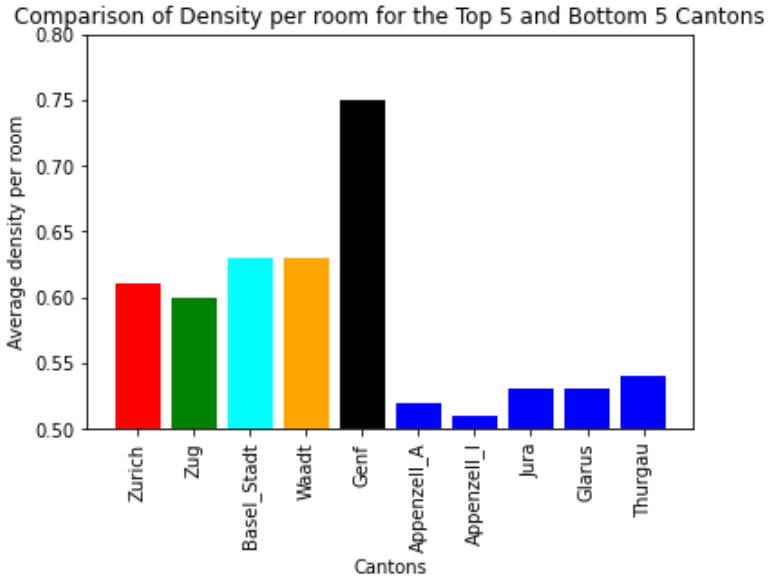

Figure 14: Average living space in square meters in a flat per person for the Top 5 and Bottom 5 Swiss Cantons for the year 2020

This higher density could actually mean two different things: either that there are not enough apartments and people start to share apartments, or that people tend to have less rooms because of the high prices in those areas.

Thus either the high density in Geneva is the result of high rent prices which leads people to live together or it is the result of the low availability of apartments regardless and the price level is only the representation of the low availability.

To check for this idea, we perform another analysis to check if the high density is actually the result of a lack of housing on the market by comparing the ratio of empty housing of the Top 5 and Bottom 5 cantons (see figure 15 below).

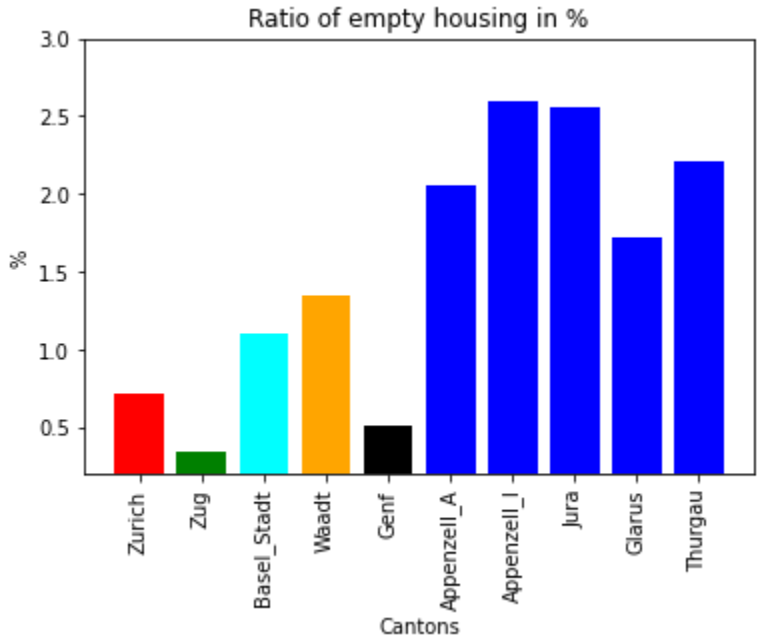

Figure 15: Ratio of empty housing for the Top 5 and Bottom 5 Swiss Cantons for the year 2020

We can see that the availability of housing in Zurich, Zug and Geneva is below 1% and on average three to four times lower than the one of the Bottom 5 Cantons. Such a low rate might have a massive impact on rent prices as people are forced to pay higher prices simply because they can not find adequate housing fast enough.

Now that we have seen an a difference in the DENSITY as well as the EMPTY HOUSING ratio, we will confirm this that there is an overall correlation between or rental price and the respective input variables:

The below figure (Figure 16) shows the negative correlation between rental prices and the amount of apartments available at any given time. The simple regression results in an R^2 of approx. 0.5746, thus a substantial price effect can be explained by this one variable alone.

Figure 16: Ratio of empty housing plotted against the rental price over time

A similar analysis of the density of people (Figure 17 below) shows a positive correlation. The simple regression results in an R^2 of approx. 0.524. Again a substantial price effect can be explained by this one variable alone.

When combining both variables into a multivariate regression model, we find an R^2 of 0.5317 which is in between the two “separate” regression models indicating that those variables are not independent of each other and therefore explaining a similar effect. This can be confirmed by running a regression between DENSITY and EMPTY HOUSING. THe result is an R^2 of 0.273, rejecting the assumption of independence of the two variables.

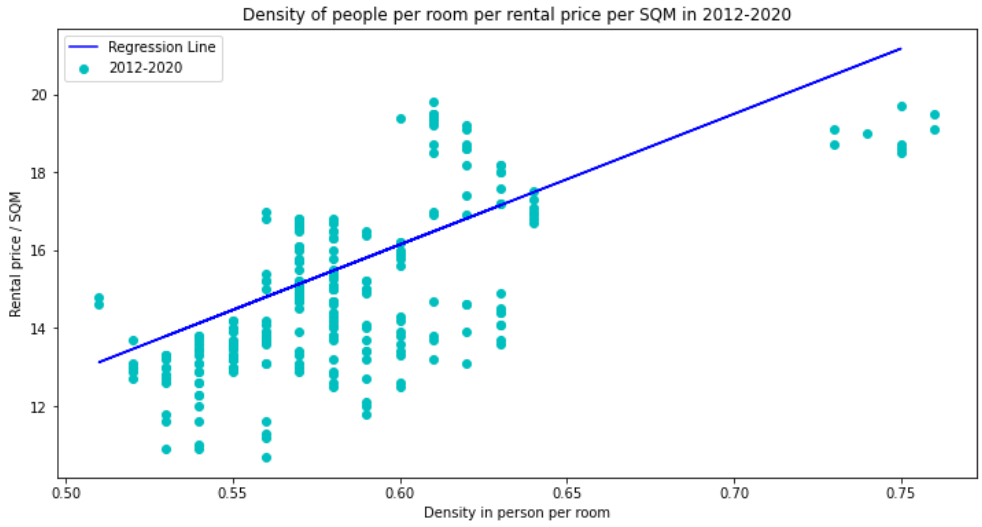

Figure 17: Ratio of density plotted against the rental price over time

As we did not control for causation, we can not state with certainty that the higher rental price is the result of an increased density or actually the reason for it.

10 Conclusions

Our very simplified model indicates that the price for rent in Switzerland is driven by the availability of housing on the market and we further believe that people react to this trend of decreasing availability by increased sharing and reducing of space of living (represented by the DENSITY function).

We believe that there are many more suitable explanatory variables to build an adequate multivariate model to better explain the rental price peaks in specific regions but additional time and effort would be required to enhance the model and take a deeper look into those variables.

Excellent candidates would be:

- the total amount of housing per canton as an indicator for supply
- the amount of social housing per canton represented by the "Genossenschaft" appartements as a mitigating factor of price increases
- the ownership quota in the different regions as an indicator of historic prices
- square meter size per apartment to control for the effect of reducing your living space when being presented with higher prices
- Income levels in specific areas (This has been ignored so far as the income data is available on another aggregation level meaning we would need to combine different cantons into bigger regions to be able to make use of such an input variable)
- the amount and the age of people per household per canton, as this can have an impact on the density
- Cantonal level tax competition that can be seen as a katalysator for population growth (and also defines the sub groups of population)
- the concentration of property in the hand of a few allowing them to set rent prices uncorrelated from the demand

A deeper analysis would also require identifying the spurious (co)relations between explanatory variables when extending the model.

It would further be very interesting to better understand the direction of the relationship (Causation) as we can argue that DENSITY can be seen as a consequence of the high prices as people tend to reduce the amount of square meters they occupy individually. At the same time, it could also be used as a proxy variable for lower income (as people tend to share space) and lower income regions are more likely to be faced with lower prices (following our supply and demand logic).

Potentially the most interesting extension of the model would be to use population growth in the future to explain current price levels. As perfect markets internalize future trends into current prices, we might be able to make an estimation of future levels of "empty houses" and see if this has an effect on the current price level. This can be modeled by using the data on growth of inhabitants per region between 2020 to 2050.

References and Bibliography

[1] Bundesamt für Statistik, Sektion Bevölkerung, Mietwohnungen, 2022, Mietwohnungen | Bundesamt für Statistik

[2] Bundesamt für Statistik, Sektion Konjunkturerhebungen, Leerwohnungen, 2022, Leerwohnungen | Bundesamt für Statistik

[3] Bundesamt für Statistik, Sektion Bevölkerung, Wohnverhältnisse, 2022, Wohnverhältnisse | Bundesamt für Statistik

[4] Frederic Bärtl et Nicolas Gaillard, Dataset for our CAS ADS M2 Project on rent prices in Switzerland, own brains, 2022, Github repository fbaertl