EDA Master

Students: Firat Saritas, Simon Stähli

Exploratory data analysis is about familiarizing yourself with the data. Let's import the dataset and see what shape we have and check a row of the dataset:

```
Shape of Dataframe: (153627, 69)
```

	Id	AreaLiving	AreaProperty	BuiltYear	FloorNumber	ForestDensityL	ForestD
147153	40521047	133.0	0.0	2017	2.0	0.341172	0

1 rows × 69 columns

This information tells us, that we have 153627 properties with 69 attributes in our dataset.

1. Describe Dataframe

Next we want to see some basic statistical details like percentile, mean, std etc. of the dataset and we want to see the types of each column

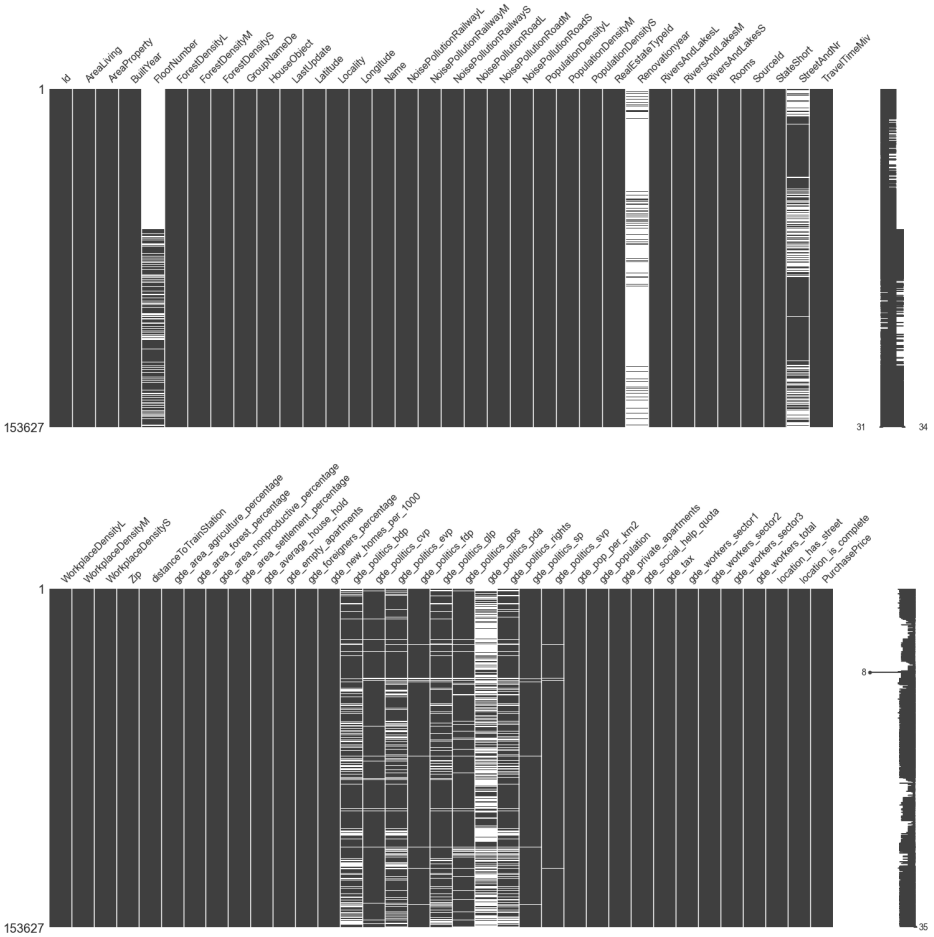
	Id	AreaLiving	AreaProperty	BuiltYear	FloorNumber	ForestDens
count	1.536270e+05	153627.000000	153627.000000	153627.000000	65932.000000	153627.00
unique	NaN	NaN	NaN	NaN	NaN	
top	NaN	NaN	NaN	NaN	NaN	
freq	NaN	NaN	NaN	NaN	NaN	
mean	2.671412e+07	141.261625	290.945532	1989.006522	12.771659	0.17
std	8.290686e+06	61.025164	490.419149	32.543506	105.121187	0.17
min	7.135329e+06	21.000000	0.000000	1800.000000	-5.000000	0.00
25%	2.430841e+07	100.000000	0.000000	1976.000000	0.000000	0.02
50%	2.638354e+07	130.000000	0.000000	1997.000000	1.000000	0.12
75%	3.112152e+07	168.000000	495.000000	2013.000000	2.000000	0.28
max	4.217594e+07	596.000000	4999.000000	2018.000000	1000.000000	0.93

```
5 example where we see the type of each column:

Id                int64
AreaLiving      float64
AreaProperty    float64
BuiltYear         int64
FloorNumber     float64
dtype: object
```

2. Missing Data

We would like to know where the missing datas are. For this, we achieve this by using the matrix of the programming library missingno, which is very useful when it comes to a quick overview over the missing values.

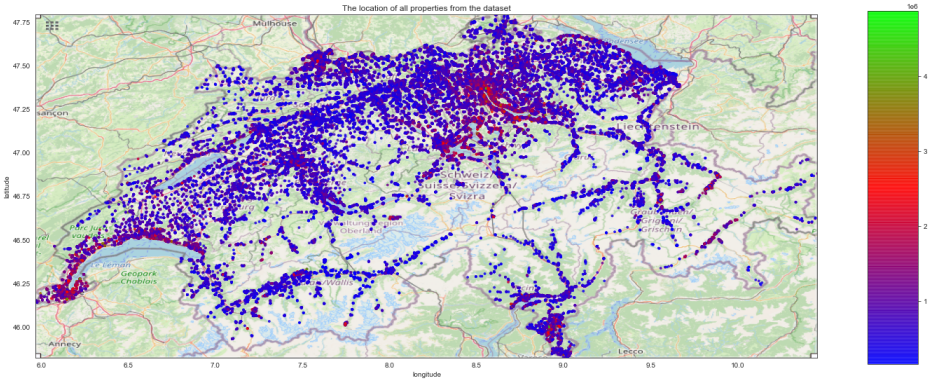


We can clearly see that missing values are very common in certain columns (MNAR).

For the column "Renovationyear" we have most of the missing data. But we can presume that the rows that have no value there have not been renovated.

3. Create a geographic plot

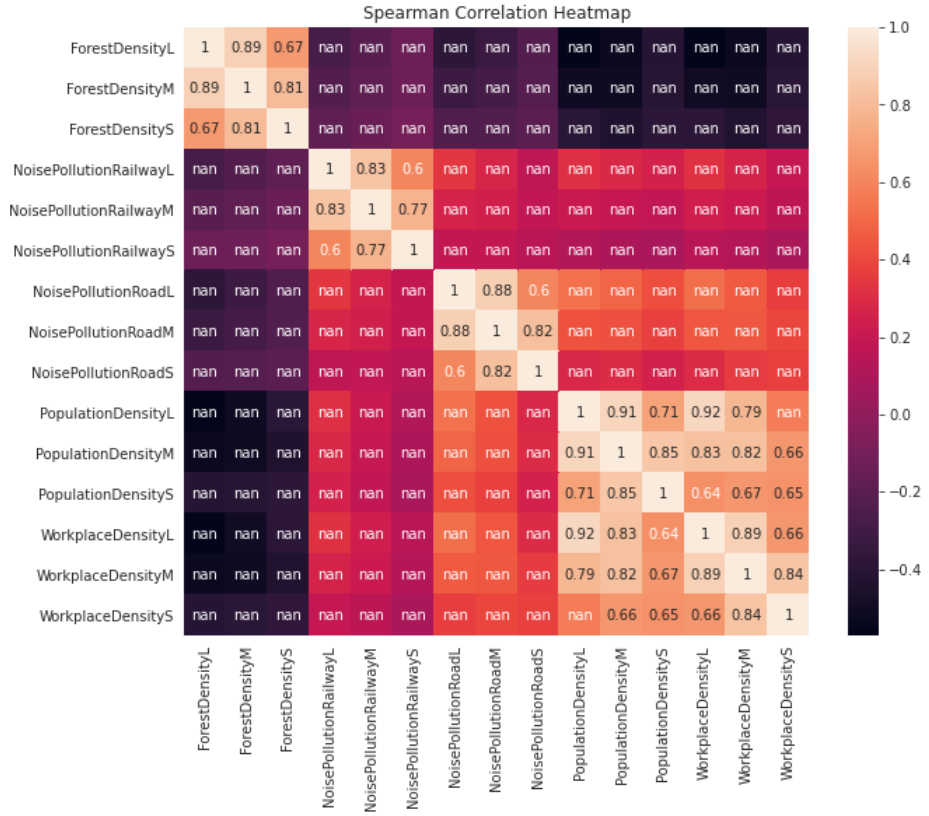
With the help of the latitude and longitude we can see the distribution of the properties on the map.



A very interesting visualization. Here you can clearly see that the prices are rising in the metropolises and at the lakes. We can also see that we have a lot more properties north of Switzerland.

3. Attributes classified by sizes S, M, L

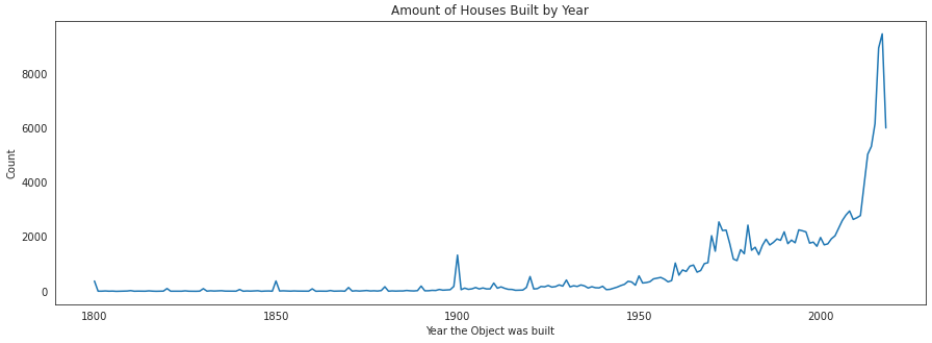
The plot below should show a relation between the different sizes S, M and L of different attributes.



Looks like there is a strong correlation with each other between these classified sizes.

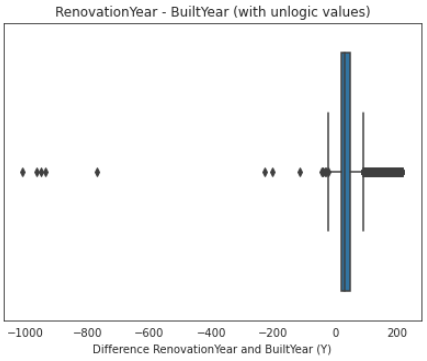
4. Built Houses by year

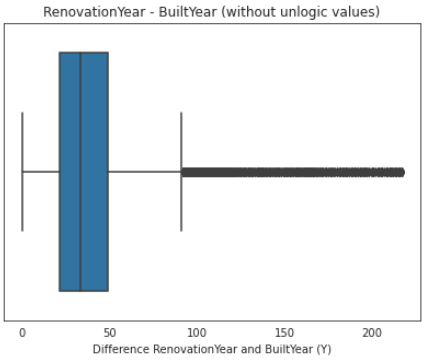
Let's have a look on the timeline when the houses were built and how much of them were built.



The plot shows the count on the y-axis and the year on the x-axis. It shows a pattern for the most of the houses which were built before 1900. The pattern consists of peaks all 10y. We presume for them that it's just an approximation of the built year. After 1900 the line fits more our expectations.

Difference Renovation Year to Built Year





These two plots show the data of a new generated data column which shows the difference between the Renovation year and the BuiltYear. Logically the RenovationYear can only took place after the BuiltYear. The left plot shows very strong outliers which have been removed for the boxplot on the right side. Additionally boh of the boxplots show upper outliers, but we estimate that these outlier could be possible i.e. a house which was built 200y ago and then renovated is possible.

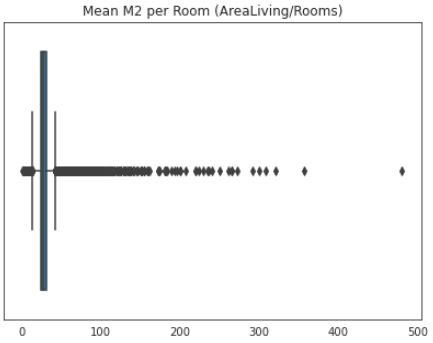
5. Crosstab SourceId and Name

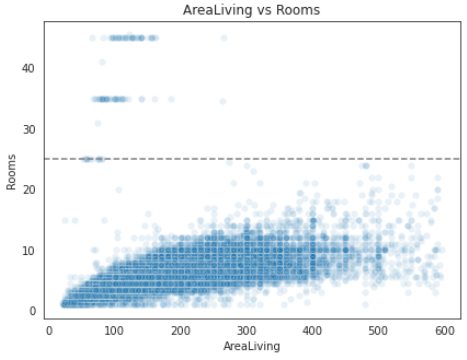
We don't know if the SourceId and the Name belong together. We investigate this by use of the crosstab function of pandas which counts the appearances of each variables in the column and puts them together in a table.

SourceId Name	10000	11000	12000	13000	14000	15000	27000	29000	30000	31000	32000
ComHistory	0	0	0	0	0	0	0	0	0	0	0
Home.ch	0	0	0	0	0	0	3262	0	0	0	0
Homegate	0	16198	0	0	0	0	0	0	0	0	0
ICasa	0	0	0	0	0	2501	0	0	0	0	0
Immoclick	0	0	0	0	0	0	0	12	0	0	0
Immoscout	0	0	20277	0	0	0	0	0	0	0	0
Immostreet	0	0	0	0	0	0	0	0	0	487	3546
Immowelt	0	0	0	0	0	0	0	0	2826	0	0
NabHome	0	0	0	0	712	0	0	0	0	0	0
Newhome	30208	0	0	0	0	0	0	0	0	0	0
Urbanhome	0	0	0	1493	0	0	0	0	0	0	0

The crosstable shows us that these values belong to each other.

6. Rooms

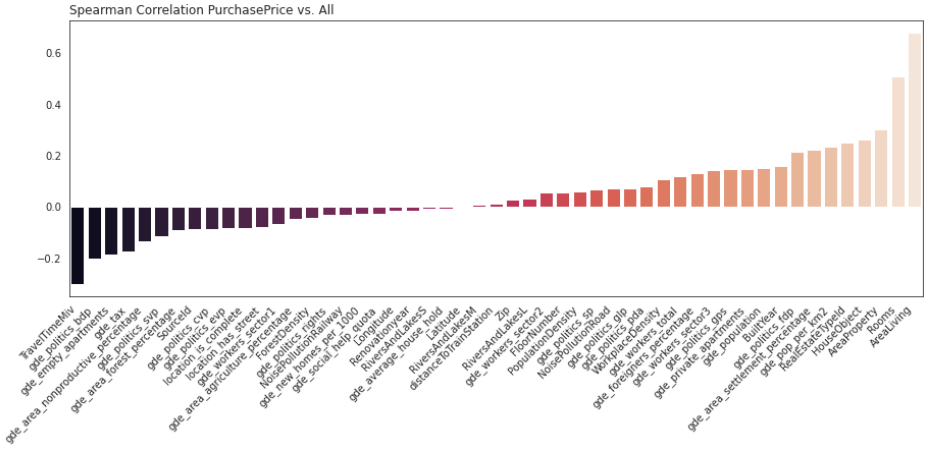




We have a lot of of outliers in the column Rooms. We wanted to investigate it with the relation to the LivingArea which is also part of the dataframe. The scatterplot shows a trend regarding to AreaLiving vs Rooms. We decided to set a filter for the amount of rooms.

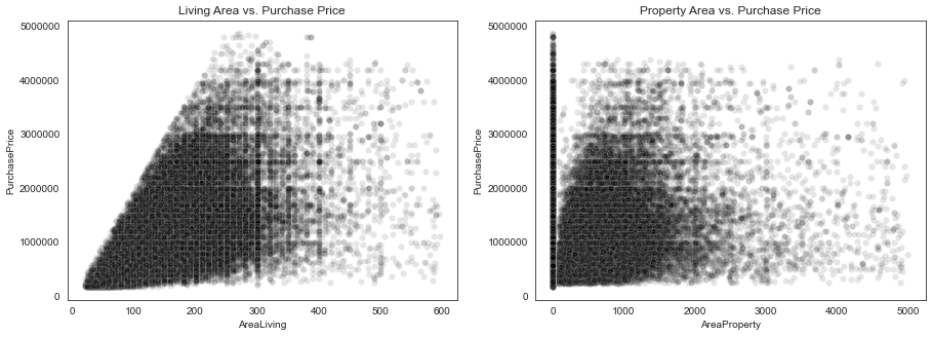
7. Spearman Correlation between Purchase Price and other Attributes

The Spearman correlation measures the relationship between two variables. It takes values from +1 (perfect positive correlation) to -1 (perfect negative correlation). And if its close to 0, there is no correlation at all.



AreaLiving has a very large correlation with PurchasePrice and latitude, for example, has one of the least correlation. TravelTimeMiv has a strong negative correlation.

Let's go over to AreaLiving and AreaProperty and compare them directly to PurchasePrice:



If we compare AreaLiving and AreaProperty with PurchasePrice, we see two interesting observations:

- In the case of AreaLiving, we see an upper limit for the price for a certain size of AreaLiving.
- at AreaProperty we see a large number of properties in all price classes with an AreaProperty of 0. We assume that this is primarily the apartments and not the Houses.

8. Distribution of the Data

Let's have a closer look on the distribution of the data and the count of upper, lower outliers and the count of the NA-Values. For this we created a funciton which is interactive to choose which attribute should be displayed.

(Because the output is an interactive visualization, this graphic is not displayed in the PDF.)

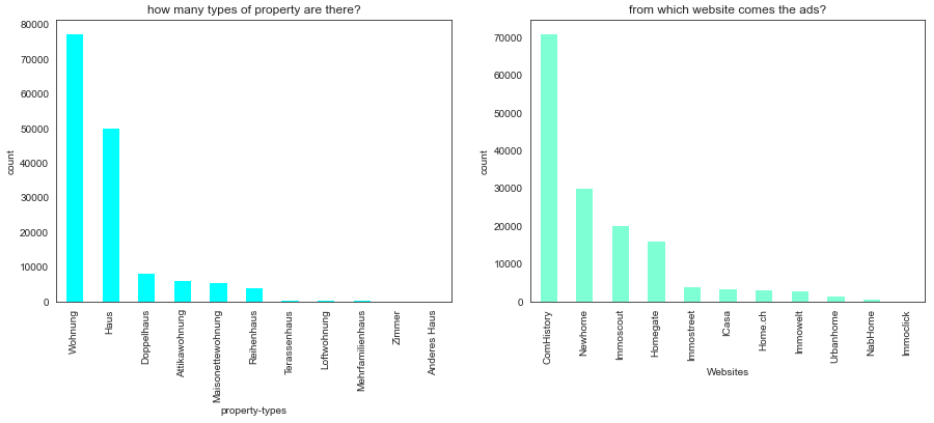
9. Whats about data which is not int or float?

Let's see how many there are and which columns this concerns:

```
Count columns: 14
```

	Id	BuiltYear	GroupNameDe	HouseObject	LastUpdate	Locality	Nam
92628	24886996	2013	Wohnung	False	2017-05-30 23:56:32	Aadorf	ComHisto
8156	17258231	1988	Haus	True	2016-12-02 19:14:13	Villeneuve FR	Immosco
130749	31610384	2003	Wohnung	False	2017-09-18 11:10:23	Frick	Newhom
97654	25392093	2013	Wohnung	False	2017-06-01 19:13:05	Oberrohrdorf	ComHisto
17780	24653320	1820	Haus	True	2017-05-30 04:03:34	Winterthur	ComHisto

How many types of property are there and rom which website comes the ads?



From these two diagrams we can see:

- There is by far a wide range of houses and apartments.
- Most properties come from the immoscout and homegate websites.
- We also see that twice as many properties are old offers.

10. Summary

Through this small analysis of the data, we were able to read out a lot of information. We have many attributes per property and among them there are missing values as well as incorrect values. these must be found and dealt with. In addition, some information is repeated and this could possibly be removed. Finally, we can say that we definitely need a processing phase for the data before we can train it with a regression model.

11. Sources

Notebooks:

Simon Staehli, S. S. (2020). EDA_Simon.ipynb [Jupyter Notebook]. Simon Staehli.

Firat Saritas, F. S. (2020). EDA_Firat.ipynb [Jupyter Notebook]- Firat Saritas.