Python project Immozil(l)a

Nathalie Nachtergaele

Jens Dedeyne

Alfiya Khabibullina

Sem Deleersnijder

https://github.com/NathNacht/immo-eliza-scraping-immozila-Cleaning-EDA.git

TABLE OF CONTENTS

01

The Subset

The subset

House/App

Due to numerous outliers within the dataset, we opted to initially segregate it into two categories: House and Apartment.

Province

Due to the substantial variations in data among the provinces, we decided to divide them.

State of building

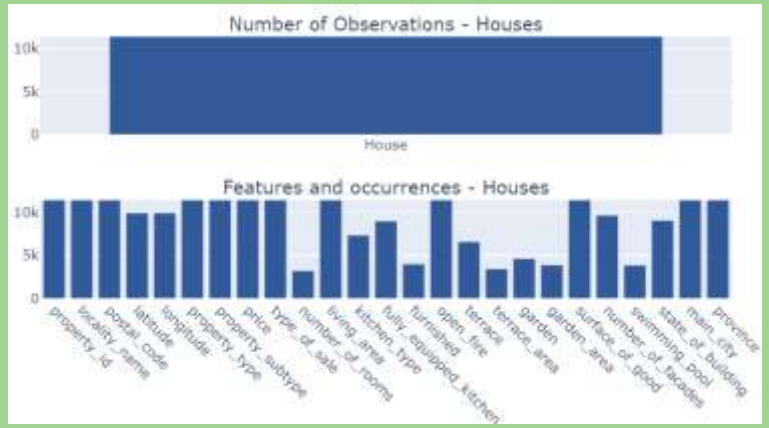
State of the building had a big impact on the price with the correlated values. So it was logical to subset this

02

The Graphs

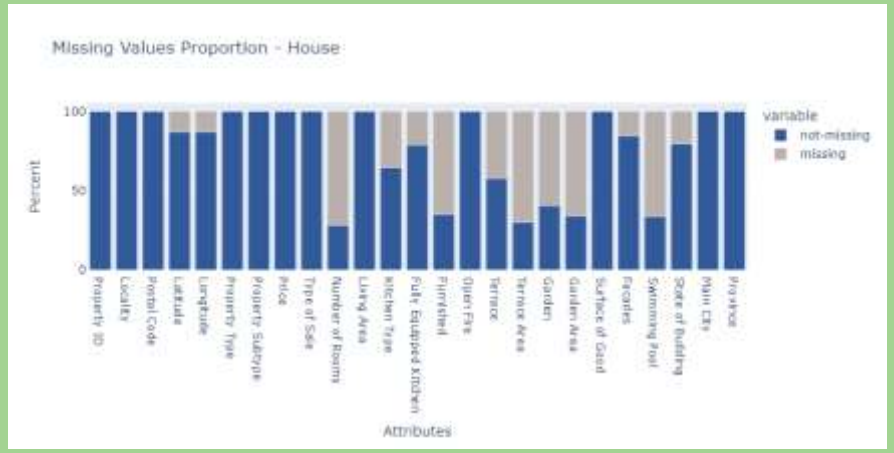
The amount of observations and features

- For the `houses` we have 11394 observations and 25 features
- For the `apartments` we have 9342 observations and 24 features (as surface of good is always empty for apartments)



The proportion of missing values per column

- Apartments has no surface of good.
- Many empty values because the form was not filled in because of non inclusion of this attribute.



What variables we would delete

Variable to remove	Reason
surface_of_good	for apartments as surface of good is always empty
property_id	as all records have a unique property_id
property_type	as for house this is always house and for apartment this is always apartement
terrace	booleans 1/0, we can deduct from terrace surface
garden	booleans 1/0, we can deduct from garden surface

These variables are most subject to outliers

The outliers are most likely to be found in quantitative, numeric data. Possible reasons for outliers are:

- **Input** is erroneous.
- **Objectively** unique, extraordinary real estate items.

In our dataset, the **price** and **various area or surface type variables (living area, garden, terrace...)** were subject to most outliers.

How did we deem values outliers?

In this project,

Interquartile Range (IQR) method was employed to count or, in particular cases, drop the outliers.

IQR is defined as the **difference** between the **75th** and **25th** percentiles of the data.

Outliers here are defined as observations that fall

below Q1 – 1.5*IQR or

above Q3 + 1.5*IQR

Let's see this on the example of the example of living area value of the apartments.

The living area of some of them was as large as 7918 m²! A manual inspection of top 10 results using Google Map Street View has shown that these large numbers are likely erroneous. Therefore, it was decided to drop them, as they influence the representation in a negative way.

The whole building looks smaller than 7819 m²...

How we handle qualitative and quantitative variables

There were three qualitative variables that could be converted into numerical values for correlation analysis. These variables included property type, property subtype, and the building's state.

Qualitative variables (18)

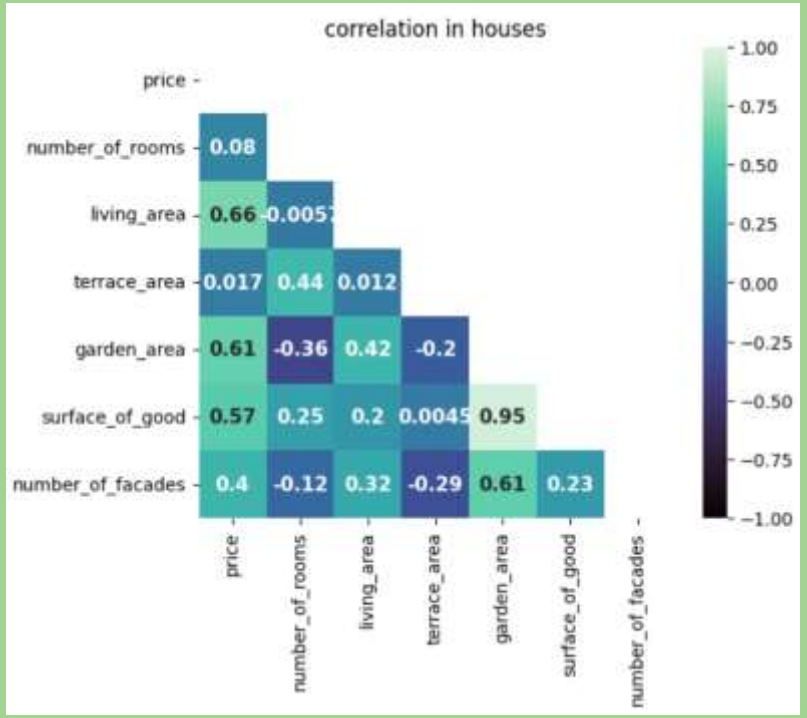
- property_id
- locality_name
- postal_code
- latitude
- longitude
- main_city
- province
- property_type
- property_subtype
- type_of_sale
- kitchen_type
- fully_equipped_kitchen
- furnished
- open_fire
- terrace
- garden
- State_of_building
- swimming_pool

Quantitative variables (7)

- price
- Number of rooms
- Living area
- Terrace area
- Garden area
- Surface of good
- Number of facade

The correlation between the variables and the price are

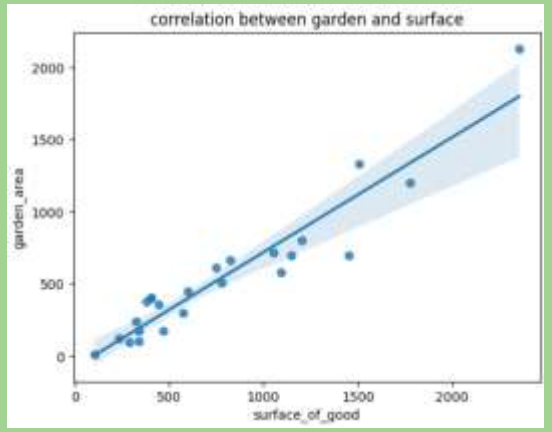
There is a correlation between price and living area, garden space, and the surface area of goods. The rationale behind these correlations is straightforward. When the price increases, it follows logically that the living space expands. Furthermore, the rise in price also leads to an increase in the land area, which consequently impacts the size of the garden. Therefore, both the surface area of goods and the garden space are influenced by changes in price.

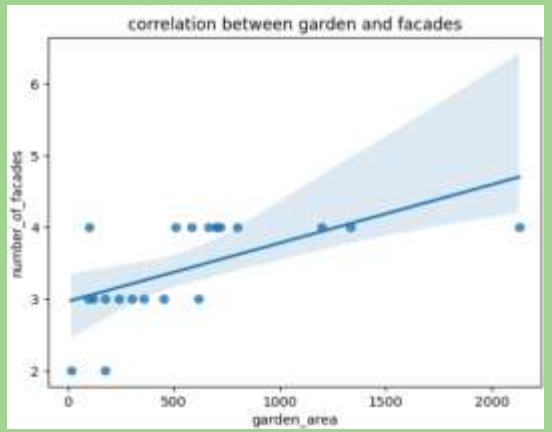


The groups of variables that are correlated together

There are correlations extending beyond price, particularly evident in the linear relationships between the surface area of goods and garden space, as well as the number of facades and garden area.

As the land surface area increases, there is a corresponding expansion in the associated garden area. Similarly, a higher number of facades, or free walls, results in an expansion of the garden space.



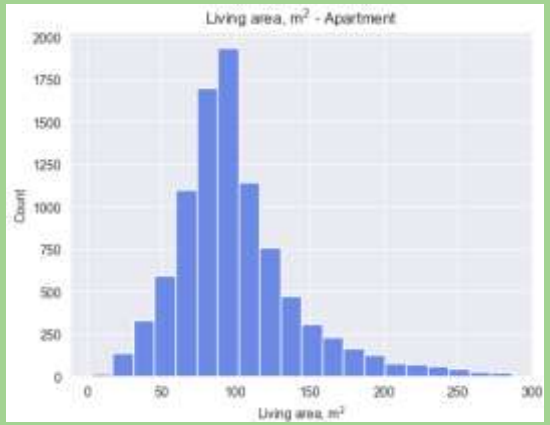


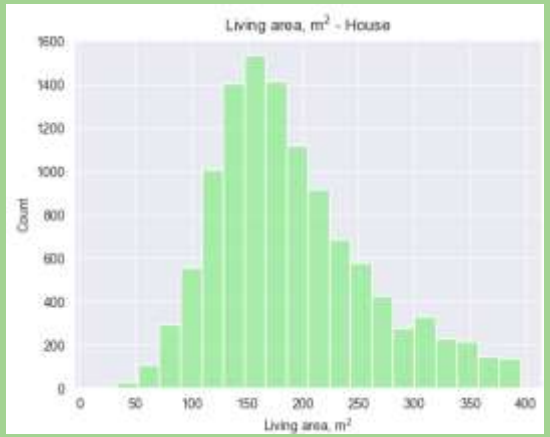
How the number of properties distributed according to their surface

The **distribution** of the living area of apartments is **right-skewed.**

This is an expected effect of the standards implemented to the housing in Belgium:

1. the **lower threshold** for the apartment area is 18 m² (24 m² for the properties that were built after 2008),
2. at the same time, there is **no upper threshold** for the living area, meaning that the extremely large apartments, e.g. penthouses will be present on the market.
3. The **most common size** of an apartment for sale is **around 100 m²**, and for the house is **around 150 m².**





Which 5 variables are most important

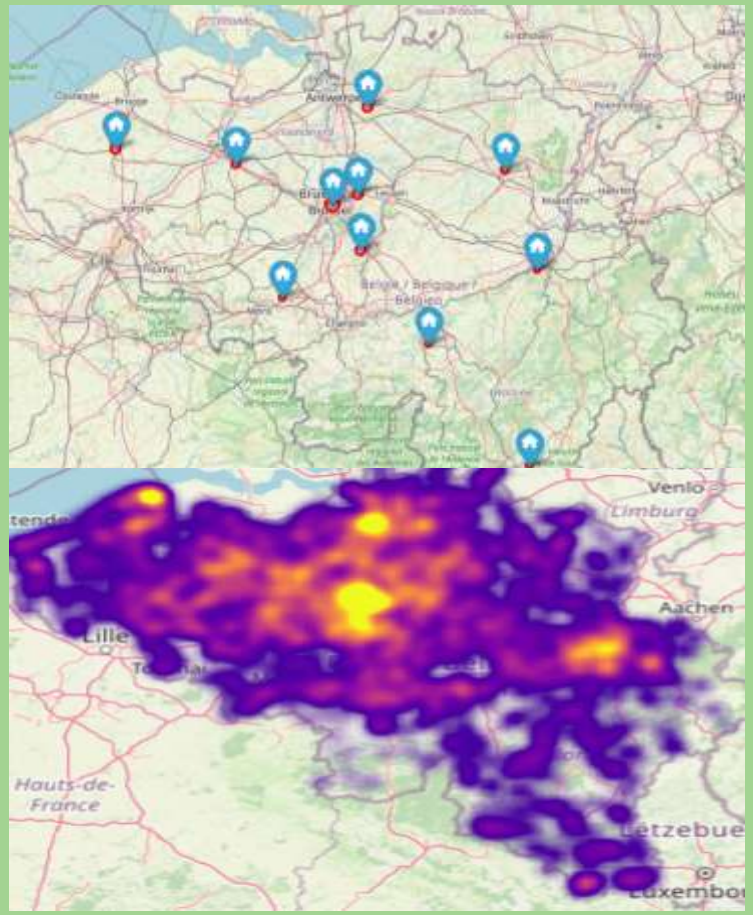
Houses	Apartments
Living area	Living area
Surface of goods	Terrace area
Province	Province
Property subtype	Property subtype
Number of facades	Garden area

Price Distribution

- The residences commanding higher market values predominantly gravitate towards prominent urban centers in Belgium.

- Notably, Brussels, Antwerp, Knokke, and Luik exhibit an elevated average pricing paradigm for residential properties.

- Conversely, more economical housing options are inclined to be positioned predominantly in the southern regions of Belgium, specifically within the Wallonian territory.



Q&A