**DATA CLEANING, ANALYSIS AND PLOTTING DESCRIPTION**

**INTRODUCTION**

In this file you will find a description of most of the steps we took in order to analyze our data.

This data has it source on immoweb, through scrapping originally around 14000 property assets.

**DATA CLEANING**

1. **Setup and dtypes changes and basic cleaning**

There were 3 libraries used: pandas, seaborn and matplotlib.

We started by making a copy of the data frame in order to keep the original one intact. From here our next step was to convert the data type of the price. This was originally a string type, that we modified to a float. This is an important step in order to make this variable quantitative.

Next step was to rename some columns names to make them more code writing friendly.

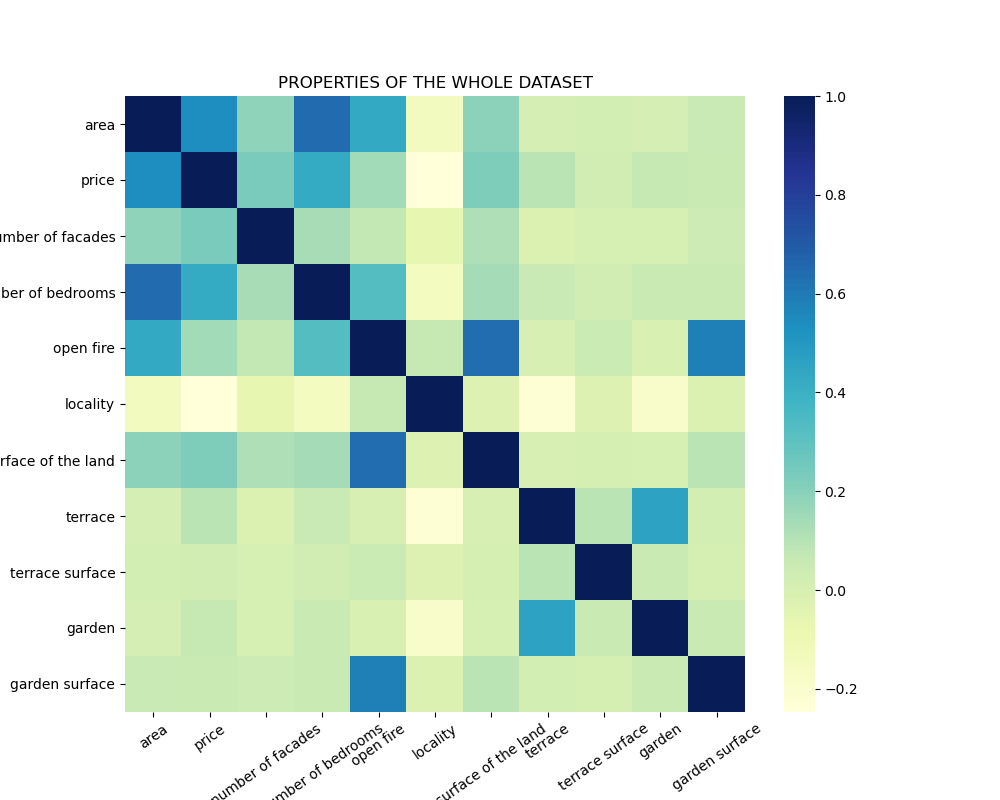
There was a column named “Unnamed” with no data inside. For this reason we dropped this column.

We looked for duplicates. To achieve so it was important to select different variables and compare if the same values showed duplicated. If so, which was the case, we deleted this duplicates.

We found that there were 2 columns with similar information. One, “type of property” contained the general information of the other, “subtype of property”. By this observance we decided to keep only the “subtype of property” since it contained more specified information.

1. **Correlation map for diagnosis and missing values**

To go forward in cleaning our data, we had to understand how deeply connected the data was with the price and what was the amount of missing values. To assess the first problem, we generated a heat map with the correlation towards price.

****

From doing so, we realized the postcode was still a quantitative element that we would have to convert into a categorical variable. So, we created 3 new columns with region, city and province. In order to assess the second problem, we printed the amount of missing values from all the variables.

Related to the missing values in area, we dropped all the houses with no area. These were around 1734 properties. The area has a strong correlation with the price, and missing values in this variable would not be desired.

From our correlation map we dropped the following columns: “terrace surface”, “surface of the land”, “open fire”, “type of property”, “fully equipped kitchen”, “furnished”, “swimming pool”, “garden”, “terrace”, “number of bedrooms”.

We wanted to treat certain features of the properties as a bonus, such as the presence of swimming pool and open fire. However, by the lack of time we did not achieve to fit this bonus in our data analysis.

In a further point of our work, we encountered about 2400 properties missing values in the column correspondent to the “state of building”. To understand the importance of this column we made a graphic with the price relating to the state of building.

Chart, histogram

Description automatically generatedIt was possible to conclude that the state of building has an impact on the price. Because of this, we assigned a new value to the state of building which specifies “not specified”. Also, the state of building is not included in a lot of houses for this reason. Lack of information from the person who created the property page.

1. **Outliers and price deviation**

Graphical user interface

Description automatically generated with medium confidenceOn our data we detected outliers in the price, surface of the garden and surface of the land. This was possible to visualize by creating a box plot, as showed bellow.

From here on, we started to think about the conformity of the price. The outliers where 4 properties with incredibly special features, such as 12 bedrooms, area of 1200m2 and 12 facades. It was easy to conclude that there were very unique properties (such as castles) , with evaluation based on many different criteria’s that we could not consider in our analysis. So, the next step was about creating a roof for the price. It was an important discussion since this project is also about creating a prediction model for prices for houses. To achieve some insight, we created a graphic to understand the distribution curve of the prices.

Chart, histogram

Description automatically generated

Here it was possible to understand that there is an undesired deviation to the right. We started reducing the price limit slowly and measuring how many properties would be lost by doing so. At a level of 750000€ and losing only 110 properties we were satisfied with compromising this way. So we dropped all this properties and kept the 750000 limit price. With this we achieved a more symmetric curve, as showed below.

Chart, line chart

Description automatically generated

1. **Adding price by square meter**

To make a more fare comparison between properties we wanted to understand the price per square meter. To do so we added a new column with this value for each property.

**DATA ANALYSIS**

We started by calculating the average, minimum and maximum price per square meter in the 3 regions of Belgium. These are:

* Bruxelles: **minimum price/m2** = 1000€; **maximum price/m2** = 11000€; **avarage price/m2** = 2447.9€
* Vlaanderen: **minimum price/m2** = 1500€; **maximum price/m2**= 9992€; **avarage price/m2** = 2037.3€
* Wallonie: **minimum price/m2** = 1300€; **maximum price/m2** = 7973€; **avarage price/m2** = 1584.6€

It is clear that Bruxelles is the most expensive, followed by Vlaanderen and finally Wallonie. This is well illustrated in the following graphic.

Chart, bar chart

Description automatically generated

Chart, histogram

Description automatically generated

With this graphic we can also conclude that villa’s are the most expensive type of property, followed by houses and finally apartments. There is an exception in the case of Vlaanderen where an apartment costs more than a house.

Again, we mention the graphic that relates the prices per region with the state of building.From this graphic we can assume that houses that are “as new”, in “good condition” and “just renovated” are the most expensive. Properties “to restore” or to be “renovated” come as the cheapest.

Finally, we wanted to go deeper and analyze the prices by city. To do so, we created a graphic with the following information.

Chart, bar chart

Description automatically generated

From this graphic what stands out is the prices in Bruxelles, being the most expensive city in Belgium, and Namur being the cheapest.