DATA CLEANING, ANALYSIS AND PLOTTING DESCRIPTION

INTROCUCTION

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 15000(?) property assets.

DESCRIPTION

LIBARIES: -Pandas

-Seaborn

-Matplotlib

COPY DATA FRAME

CHANGE DTYPE for price, from str, to float

Counting non values in data frame:

- area – very important we decided to drop 1734

-

-

-

-

DROPING DUPLICATES:

We used 2 variables

REGIONS:

We created to columns with municipality and city. A postecode is a quantitative data, and was being avaluated as a number and not as a region. In order to get a relation to regions, we did the correspondence of the postecode with their related areas.

Correlation map:

High correlation:

STATE OF THE BUILLDING:

State of the building: 2400 missing state of the building. We decided to fill this missing values with “not described” since this values correspond to about 25% of our data. Normally the state of the building is not specified

We dropped properties with the no area.

ADDING PRICE BY SQR METER

Dropping outliers: price 9500000!!! To understand if we should delete is, we got some features of this house, such as area 1100 m2, type of property is house

OUTLIERS:

They are very special properties that we have to take into consideration different parameters very luxuous assets, we decided to exclude them. For this we exclude all castles