Cleaning Data and EDA

Aldo Narvaez Cao Romero

May, 2021

The first five rows of the used data can be observed in figure (1), which is the info of house sales. 5000 rows and 16 columns

index	MLS	sold_price	zipcode	longitude	latitude	lot_acres	taxes	year_built	bedrooms	bathrooms	sqrt_ft	garage	kitchen_features	fireplaces	floor_covering	HOA
0	21530491	5300000.0	85637	-1.103.782	31.356.362	2154.0	5272.0	1941	13	10	10500	0	Dishwasher, Freezer, Refrigerator, Oven	6.0	Mexican Tile, Wood	0
1	21529082	4200000.0	85646	-111.045.371	31.594.213	1707.0	10422.36	1997	2	2	7300	0	Dishwasher, Garbage Disposal	5.0	Natural Stone, Other	0
2	3054672	4200000.0	85646	-111.040.707	31.594.844	1707.0	10482.0	1997	2	3	None	None	Dishwasher, Garbage Disposal, Refrigerator	5.0	Natural Stone, Other: Rock	None
3	21919321	4500000.0	85646	-111.035.925	31.645.878	636.67	8418.58	1930	7	5	9019	4	Dishwasher, Double Sink, Pantry: Butler, Refrigerator	4.0	Ceramic Tile, Laminate, Wood	None
4	21306357	3411450.0	85750	-110.813.768	32.285.162	3.21	15393.0	1995	4	6	6396	3	Dishwasher, Garbage Disposal, Refrigerator, Microwave, Oven	5.0	Carpet, Concrete	55

Figure: First five rows of the used data to be cleansed

```
MLS                int64
sold_price       float64
zipcode            int64
longitude         object
latitude          object
lot_acres        float64
taxes            float64
year_built         int64
bedrooms           int64
bathrooms         object
sqrt_ft           object
garage            object
kitchen_features  object
fireplaces       float64
floor_covering    object
HOA               object
dtype: object
```

Figure: Types of variable of every column in the data

Searching for None values

- This value can only be assigned to categorical variables.
- Apply an algorithm to search for the None values in each categorical column.

```
bathrooms:
6
sqrt_ft:
56
garage:
7
kitchen_features:
33
floor_covering:
1
HOA:
562
```

Figure: Categorical variables with their amount of None values

- Locate the variables that are supposed to be numerical and, in every None value assign a zero
- Transform the whole column into a float or integer.
- Observe for which variable it is ok to have zero values and work with the ones that must have a value different from zero.
- Extract some statistical information to know which value would be the best to put in the zero cases.

In the case of year_built, lot_acres, sqrt_ft and probably HOA should not supposed to have zero value, in contrast with bathrooms and garage.

Since for HOA there wera several values, i.e. the mode, with value zero, we use this for the zero values. We used the mode as well for the year_built variable.

For the sqrt_ft variable we compute the median, the mean and the mode. And we got the results 3715, 3524, 3674. We use the mean in this case

Now, we got rid off the null values. So, by using the .isnull().sum() method we compute the number of null values in our data frame.In our case were lot_acres and fireplaces with 10 and 25 null (or NaN) values respectively.

We dropped them following the 5% criteria and the whole the MLS column, due to relevance.

Correlation

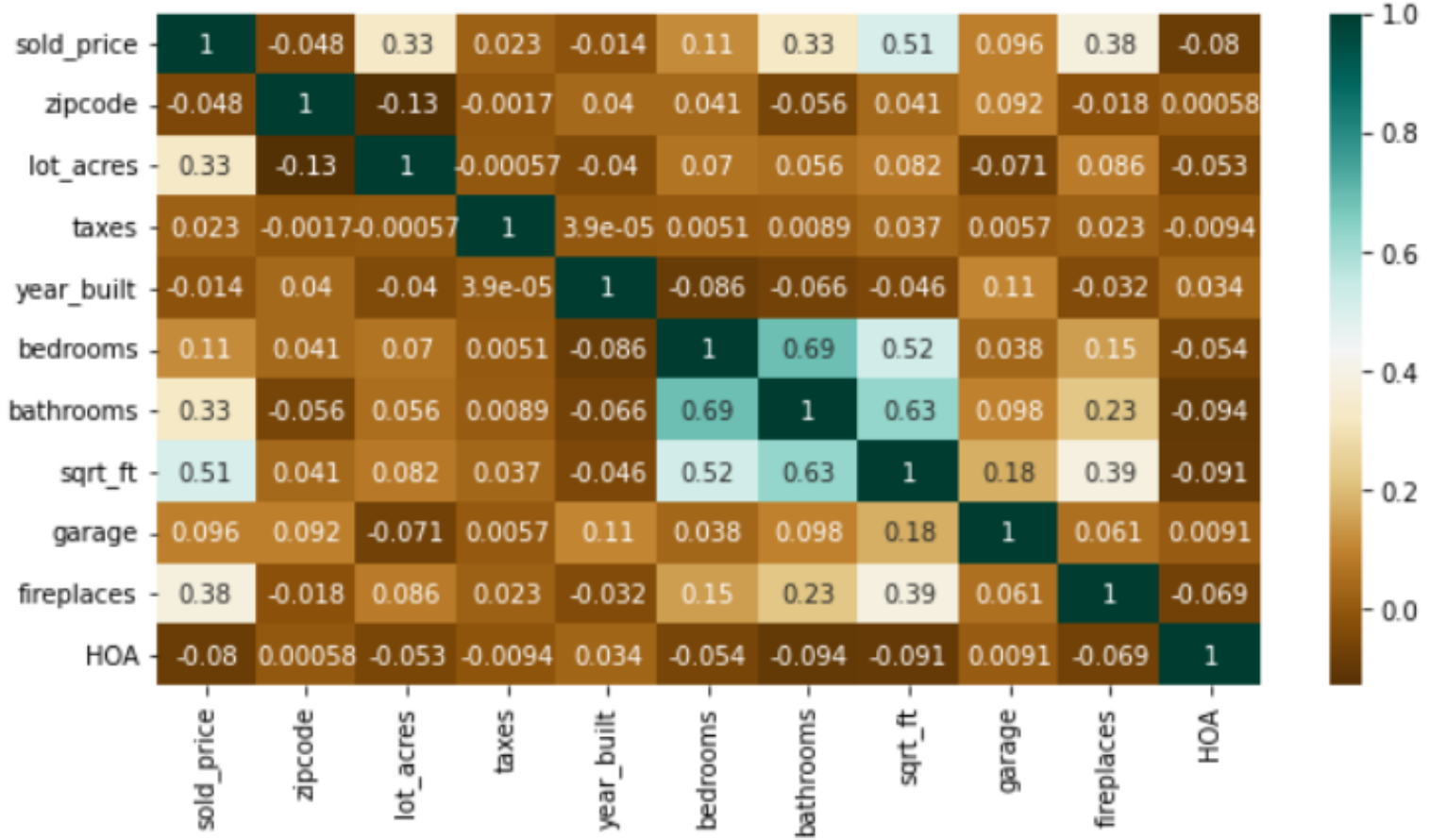

Figure: Heatmap of the correlation matrix showing the correlation among variables.

Outliers

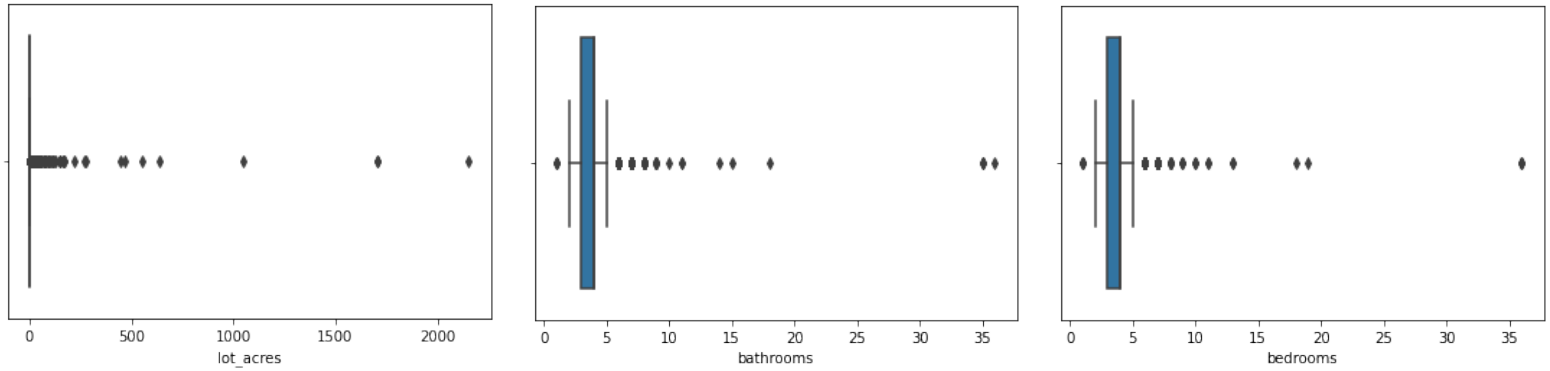

Figure: Box plots made to identify the outliers for the variables lot_acres, bathrooms and bedrooms.

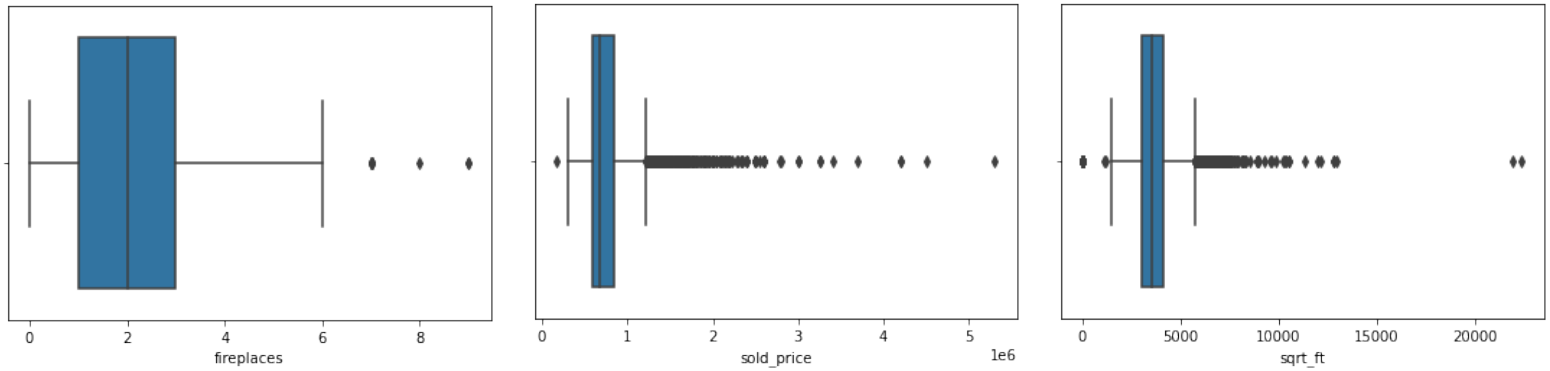

Figure: Box plots made to identify the outliers for the variables sold_price, sqrt_ft and fireplaces.

In the case of lot_acres we use

$$\text{data\_frame} = \text{data\_frame}[\tilde{}(\text{data\_frame}[\text{"lot\_acres"}] > 500)]$$

$$\text{data\_frame} = \text{data\_frame}[\tilde{}(\text{data\_frame}[\text{"lot\_acres"}] == 0)]$$

Once this was made we compute the shape of the data and it was (4916,15)

Encoding kitchen features nd floor covering

kitchen_features	fireplaces	floor_covering	HOA	kitchen coder	floor coder
Dishwasher, Freezer, Refrigerator, Oven	6.0	Mexican Tile, Wood	0	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 1.0, 0.0, ...	[0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, ...
Dishwasher, Garbage Disposal	5.0	Natural Stone, Other	0	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 1.0, 0.0, ...	[0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, ...
					[0.0

Figure: Result of the encoding kitchen_features and floor_covering.