Raw House Data Cleanse

German E. Baltazar Reyes

05/23/2022

Dataset Description

5,000 observations regarding the qualities and cost of different houses

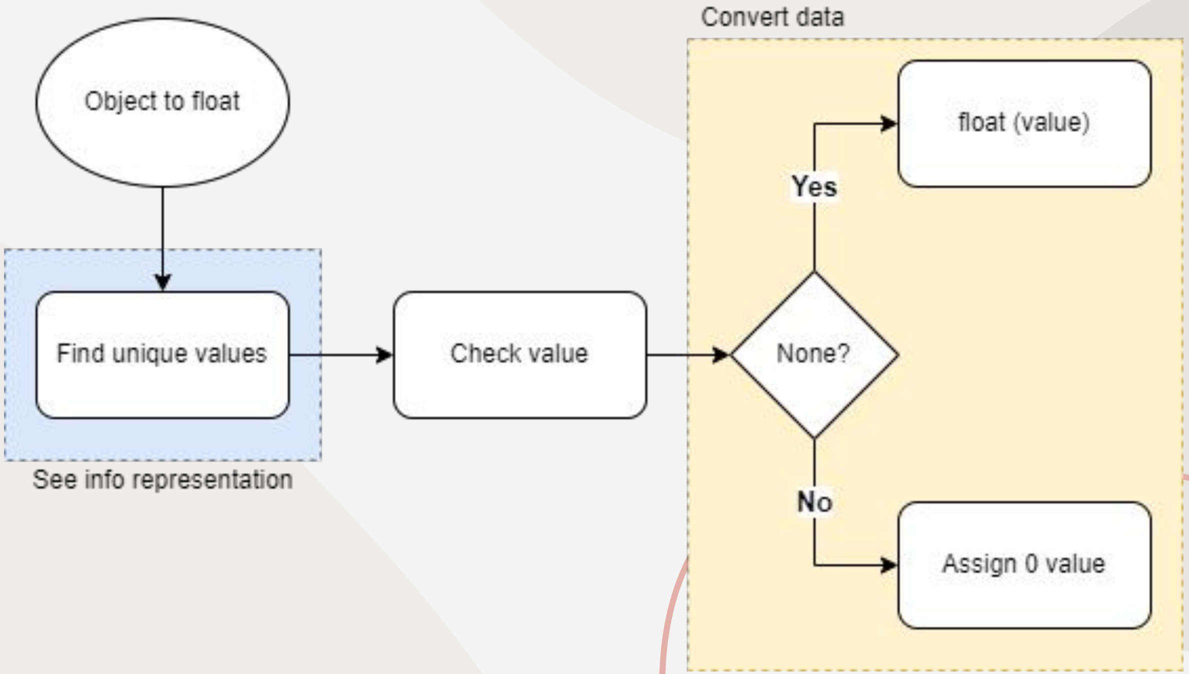
	Int	Float	Categorical
MLS	X		
sold_price		X	
zipcode	X		
longitude		X	
latitude		X	
lot_acres		X	
taxes		X	
year_built	X		

	Int	Float	Categorical
bedrooms	X		
bathrooms			X
sqrt_ft			X
garage			X
kitchen_features			X
fireplaces		X	
floor_covering			X
HOA			X

Data Type Conversion

Procedure for *bathrooms, sqrt_ft, garage,* and *HOA*

It was considered since there are other features with null values.



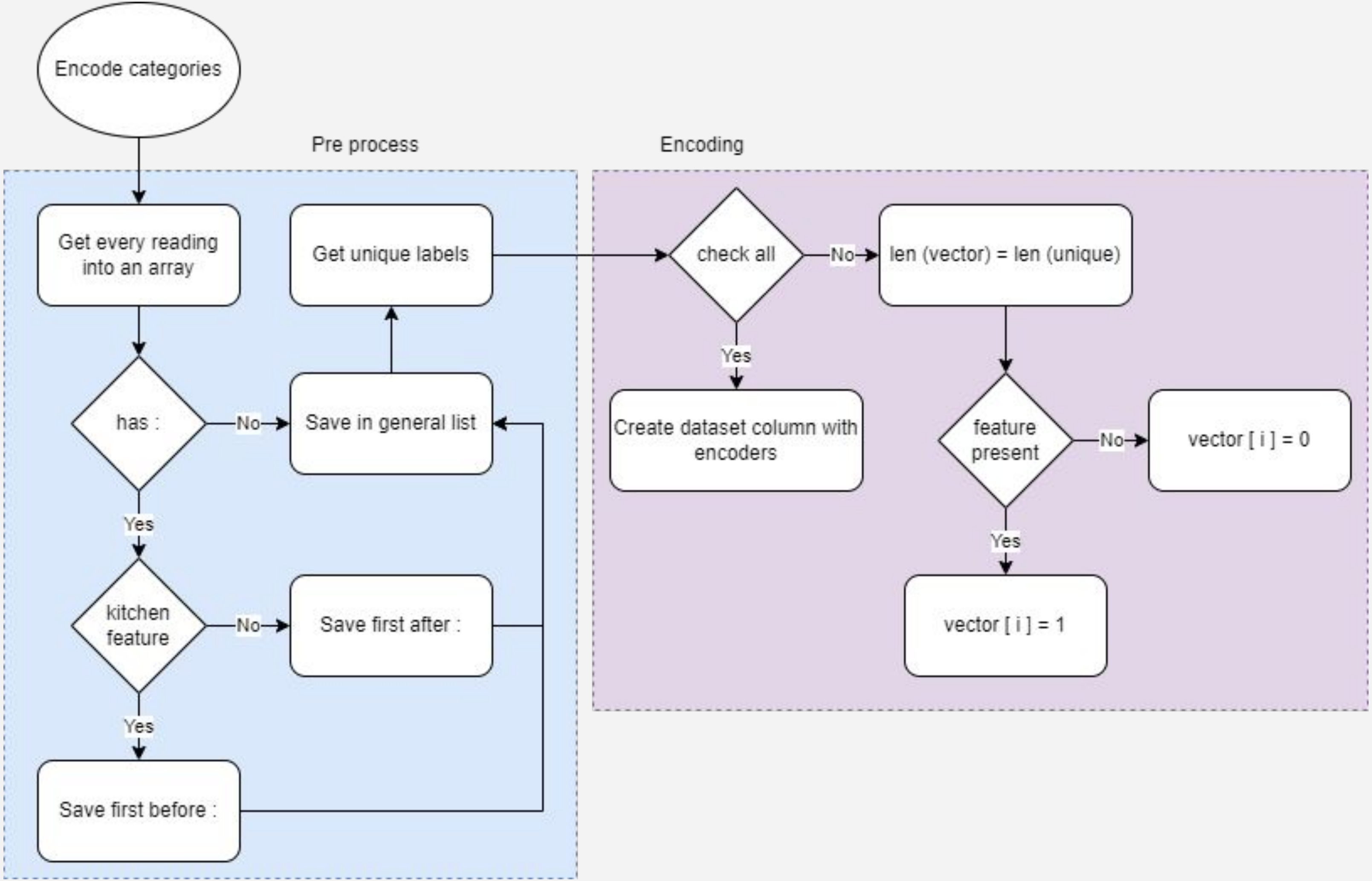
Encoding Categorical Features

Procedure made for *kitchen_features* and *flor_covering*.

Used multi-class encoding to get a single vector with every characteristic present for each house.

Created two new columns in the dataset:

- *kitchen_vectors* = **55-D vector**
- *floor_vectors* = **87-D vector**

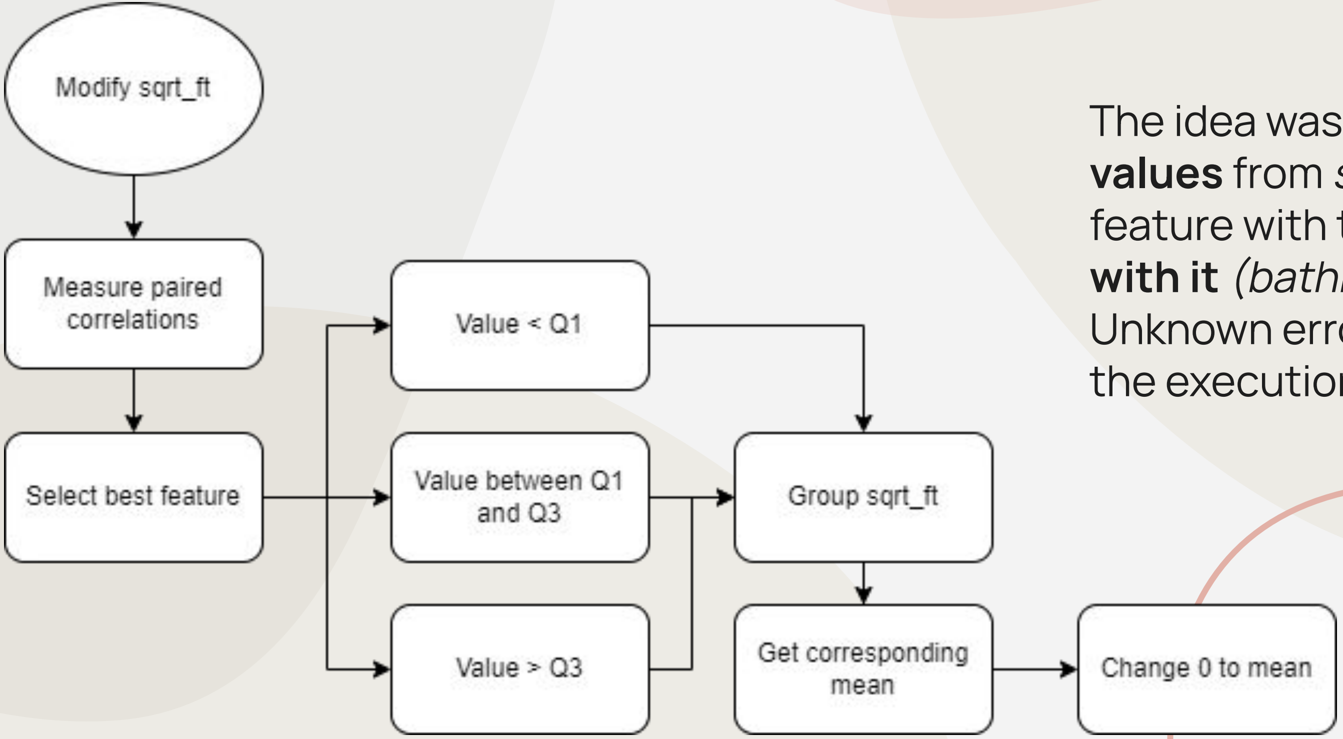

Encode categories
Pre process
Encoding
Get every reading into an array
Get unique labels
has :
No
Save in general list
Yes
kitchen feature
No
Save first after :
Yes
Save first before :
check all
No
len (vector) = len (unique)
Yes
Create dataset column with encoders
feature present
No
vector [i] = 0
Yes
vector [i] = 1

NaN and Outliers

Used EDA and boxplots to determine outliers

- Deleted null values from *lot_acres* and *fireplaces*
- Deleted column *MLS* (every value was unique)
- Used IQR method for deleting outliers in *longitude, latitude* and *fireplaces*
- Deleted *taxes* with 0 value and bigger than its Q3 score
- Deleted *sqrt_ft* with 0 value
- Deleted *year_built* before the 1800s
- Deleted *sold_price* bigger than 3M
- Kept *zipcodes* between 85400 and 86000
- Deleted *lot_acres* bigger than 500

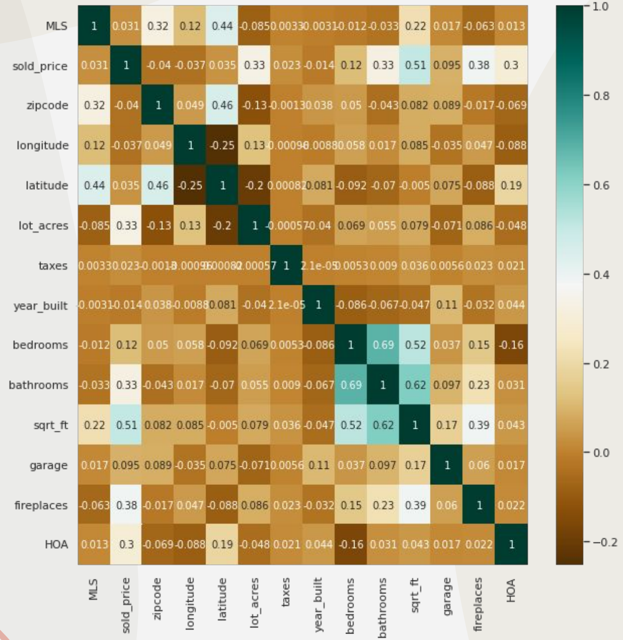
Failed Implementations

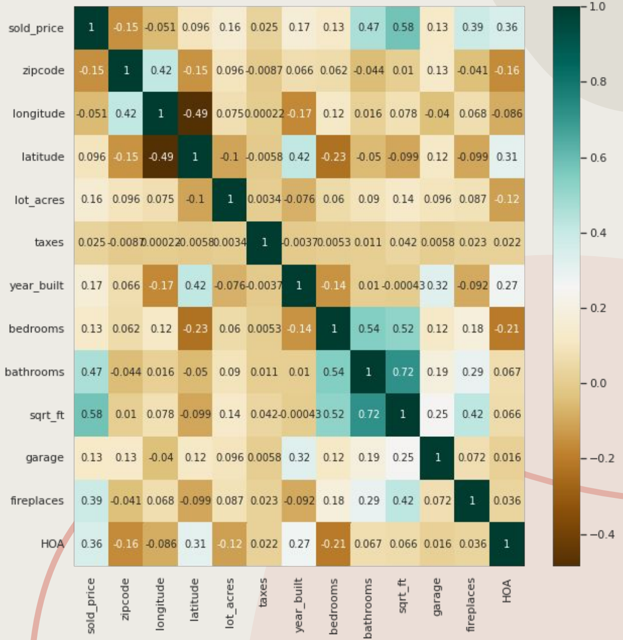


The idea was to **modify the zero values** from *sqrt_ft* based on the feature with the best **correlation with it** *(bathrooms)*
Unknown errors appeared during the execution of the code.

Conclusions

- Deleted a total of 373 observations (7.46%) from the original dataset
- Included encoded representation of the two categorical variables
- A better correlation of features can be observed after the cleanse of data





Thank You!

Any Questions?