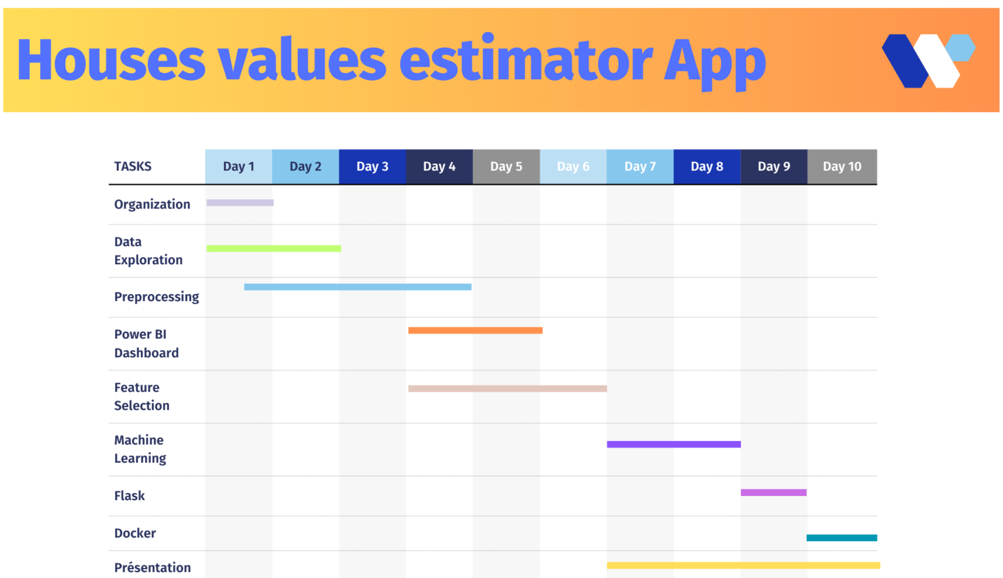
Présentation_notebook

May 6, 2024

1 Apartment-Hunter La Plateforme Project

1.1 Roadmap



1.2 Data Exploration

1.2.1 Choosing a Dataset

Madrid Pros :

- A hefty number of features (58)
- Some features are available only in this dataset (has_parking , rent_price = cheating)

Madrid Cons :

- Questionable dataset quality
- ex: 24% of listed houses have pools while only 8% have gardens.
- Only 26/58 features contain under 30% missing values.
- Both features mentioned earlier, has_parking and rent_price are empty.
- Some useful features in the other dataset aren't available in this one.

kc Pros :

- Overall the dataset is of better quality
- No missing data, very few duplicates
- About as many exploitable values as in the Madrid Dataset

kc Cons :

- Some useful features in the other dataset aren't available in this one.
- Has probably already been cleaned by someone else, which mean possible human errors.

The Decision **For all the reasons listed earlier, we will be choosing the kc dataset, mainly because he has a very high Trust index compared to the Madrid one.**

1.3 Data Cleaning

1.3.1 Duplicates

We do observe a small amount of duplicates in which only the price of the house change, we end up keeping the last sells.

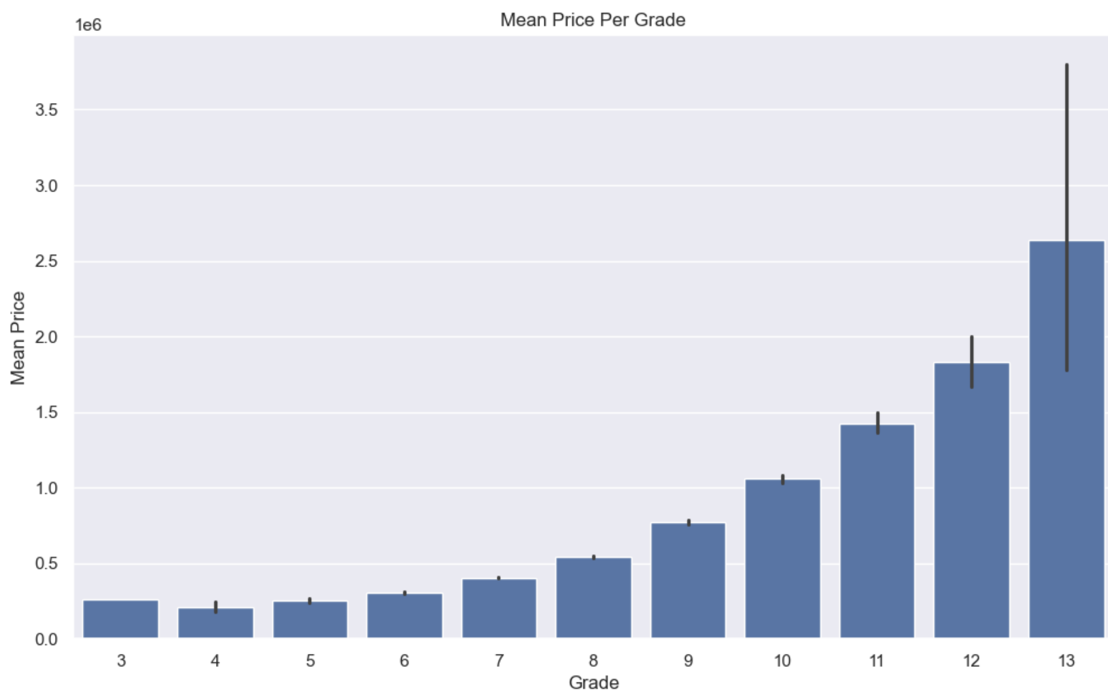
1.3.2 Categorical Features

For most of those features, we ends up dropping classes at either ends of their spectrum.

1.3.3 Grade

We end up dropping Grade classes 3-4 of and 12-13 because of their low representation, and questionable distribution in the higher end ones.

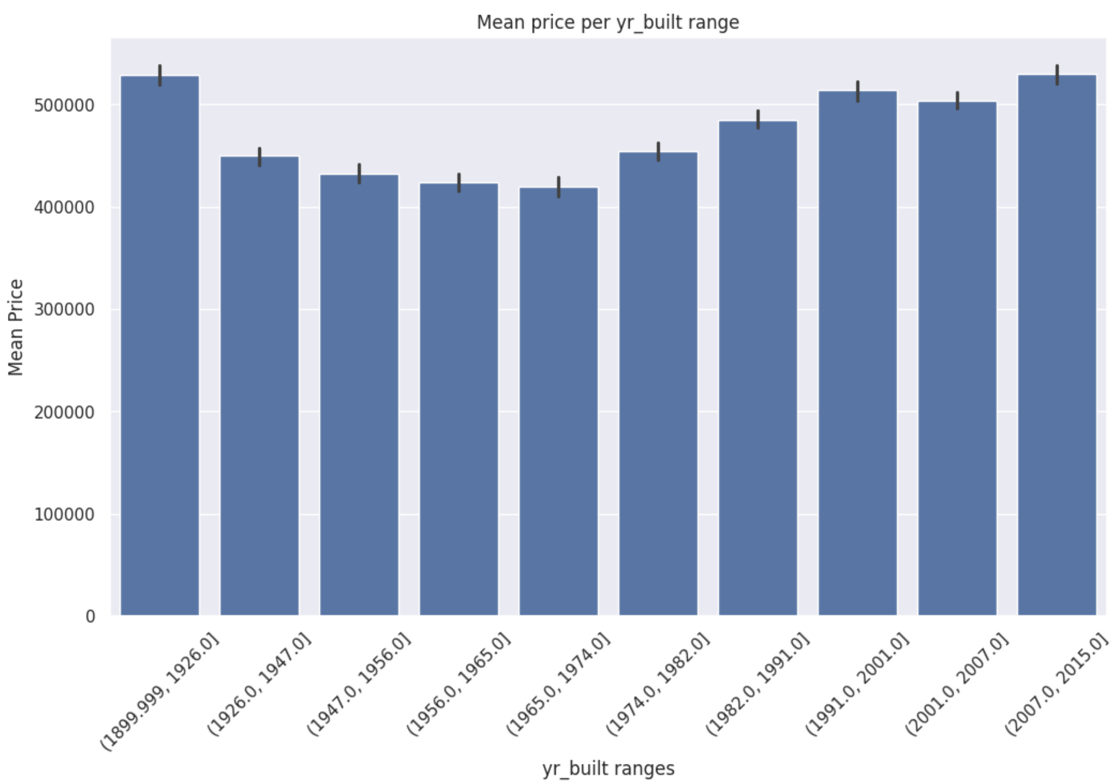
in the class "13" the range of price can go as far as 2 millions, which could disturb our predictions.



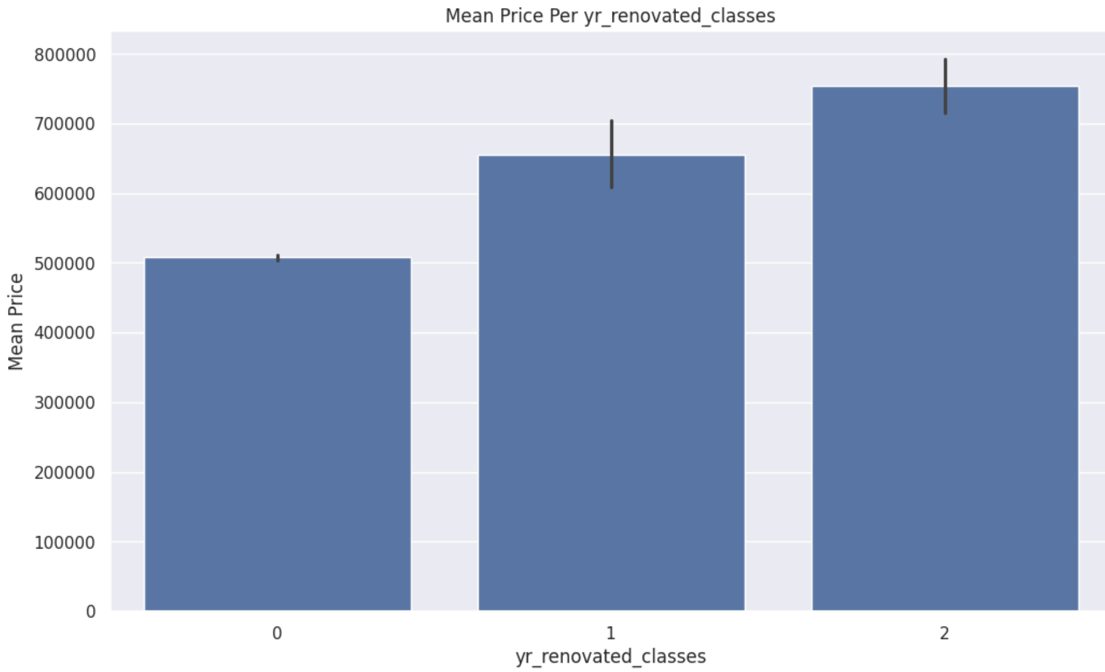
1.3.4 Numerical Features

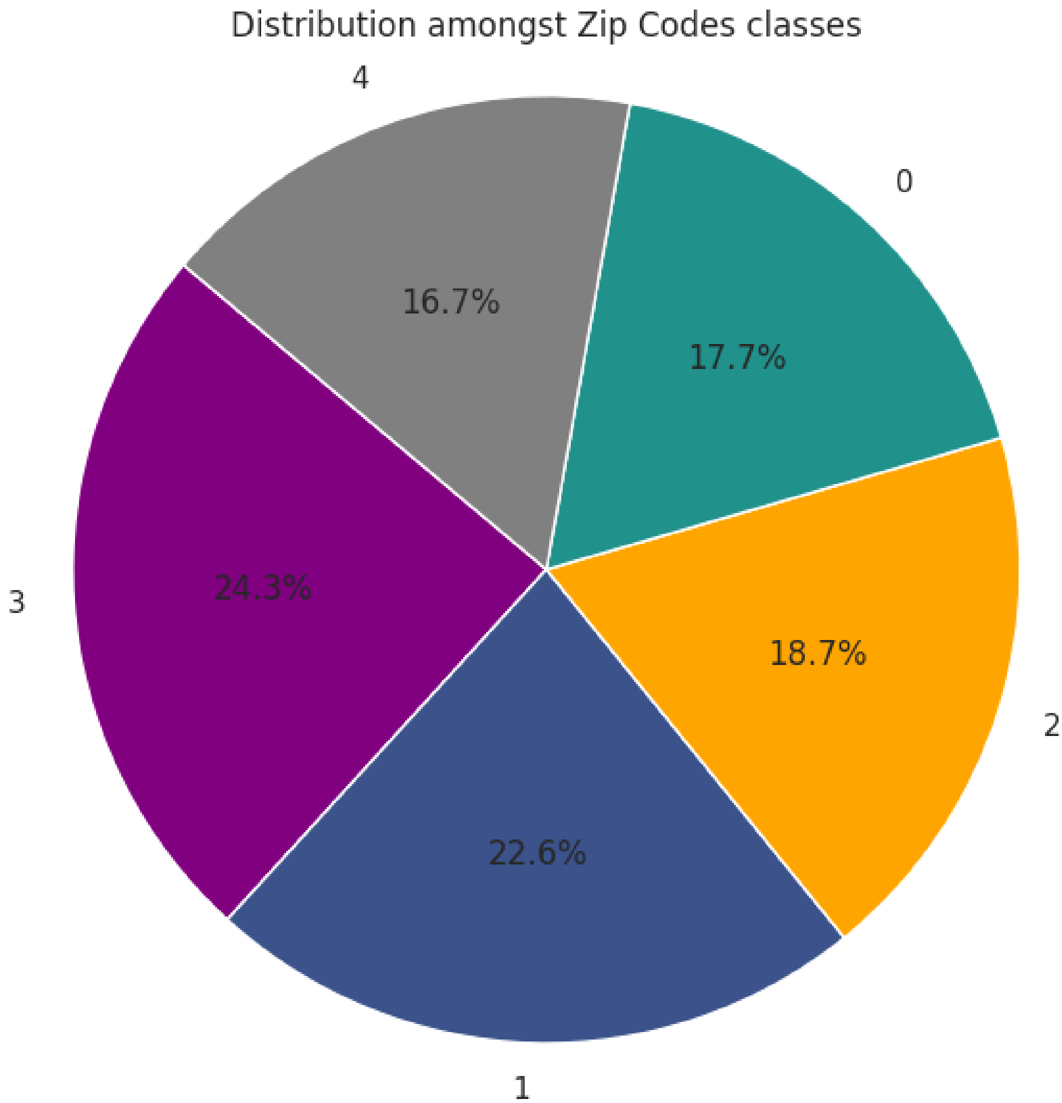
Most of these features are centered around Square footing (sqft) and have outliers values.

1.3.5 Feature Engineering

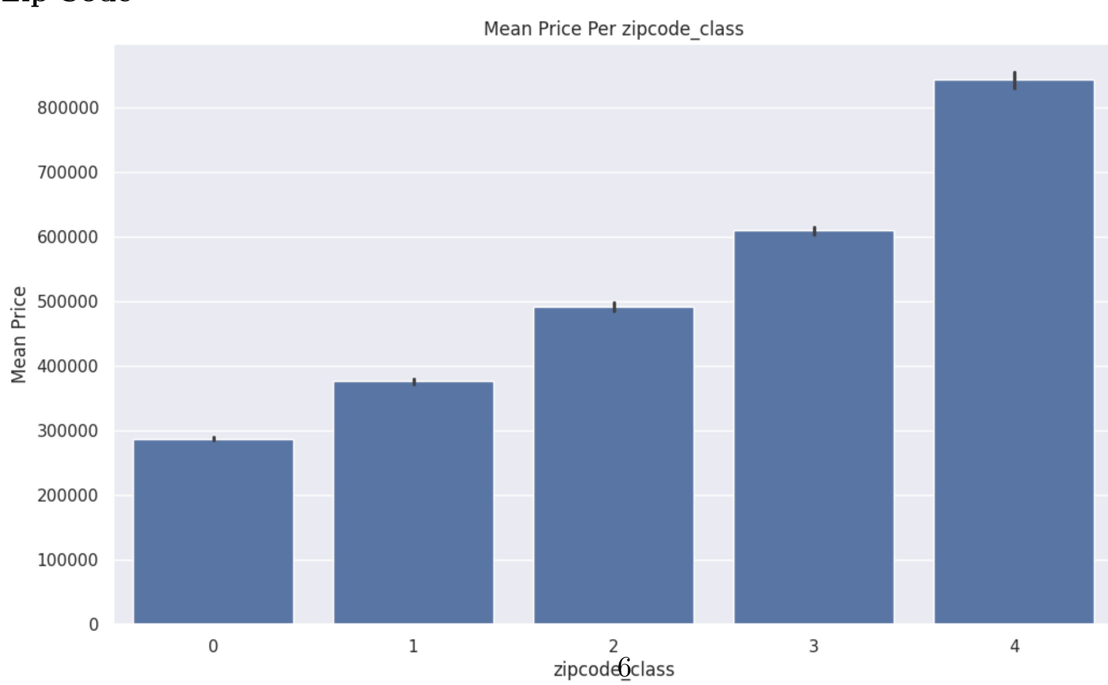


Year built After making some research we found out that houses really starts loosing values after not being renovated for the last 10-15 years so we'll be taking a wider margin of the last 20 years and only make 2 classes over those 771 values.

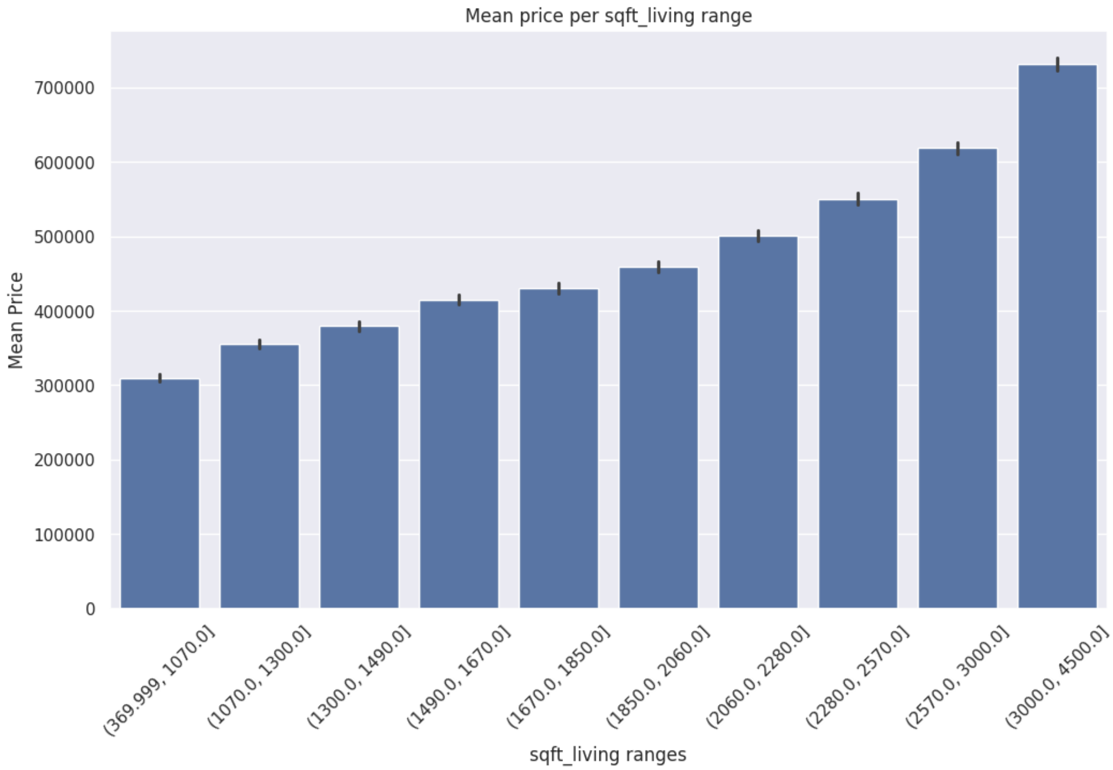



Distribution amongst Zip Codes classes
4
0
16.7%
17.7%
3
24.3%
18.7%
2
22.6%
1

Zip Code

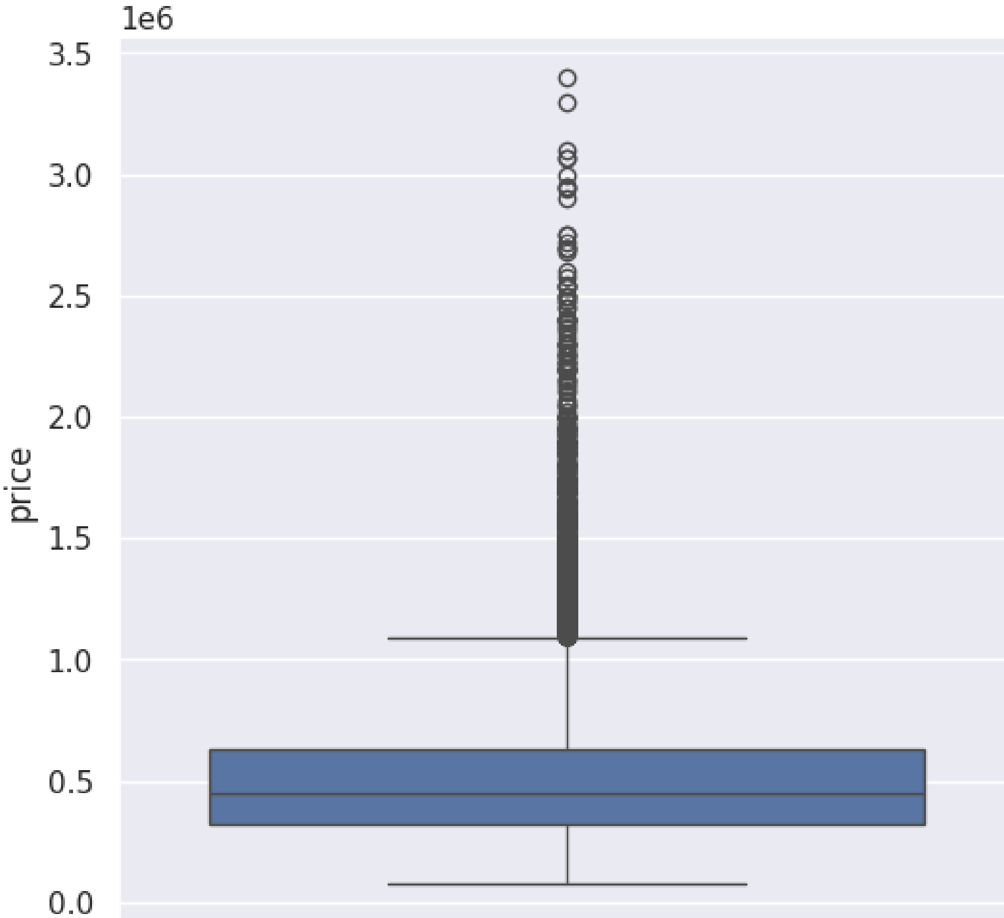

Mean Price Per zipcode_class
800000
700000
600000
500000
400000
300000
200000
100000
0
Mean Price
0
1
2
3
4
zipcode_class

It seems to have had a positive effect on the potential usefulness of the zipcode feature.



sqft_living_range

Price Outliers $(IQR = Q1 - Q2)$ $(Q1 - 1.5 * IQR - Q3 + 1.5 * IQR)$



1.4 Stay up to date with technology - Regression

1.4.1 Linear Regression

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data.

1.4.2 Decision Tree Regressor

The decision trees is used to fit a sine curve with addition noisy observation. As a result, it learns local linear regressions approximating the sine curve.

1.4.3 Ridge Regression (L2 regularization)

Ridge regression is a model-tuning method that is used to analyze any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values.

The cost function for ridge regression: $\min_{\theta}\left(||Y - X\theta||^2 + \lambda||\theta||\right)$

1.4.4 Evaluating the model's performance:

RMSE Definition :

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

It measures the average deviation between predicted and observed values, emphasizing larger errors due to the squaring process.

1.5 Feature Selection

`Eye Test for feature selection`

Probably won't be relevant

- Day
- floors
- sqft_lot
- sqft_lot15
- condition
- yr_built

Won't be given a chance

- Year
- Month

Explore options `(VarianceThreshold, SelectKBest, Boruta, Forward feature selection, VIF)` We are starting with the 11 remaining features

Most promising:

- sqft_living
- sqft_above
- grade
- yr_renovated_classes
- zipcode_class

relevant: - bathrooms - sqft_basement_class - sqft_living15

Possibly relevant: - bedrooms - waterfront - view

`VarianceThreshold` We choosed a pretty low threshold as we want to filter our remaining features with multiple tools.

Was dropped - **yr_renovated_classes** - **waterfront**

`SelectKBest` Out of our 9 remaining features, we choose to keep the **k=6** best here, price included

Was dropped - **sqft_basement_class** - **bedrooms** - **view**

`Boruta & Forward feature selection` All Good !

Multicolinearity

`VIF` All values get a pass again.

In another VIF test sqft_above and sqft_living had VIF Scores of **above 60**.

Common sense and our earlier tests established sqft_living to be our most valuable feature.

We end up dropping `sqft_above`

1.6 Machine Learning

1.6.1 With a Single feature

Model	Baseline	Linear	Decision Tree R	Ridge
RMSE	161756	160665.2	163420	`159645`
R2	X	`0.37`	0.4067	0.3691

1.6.2 With multiple features and Grid Search

Model	Baseline	Linear	Decision Tree R	Ridge	ElasticNet	SVR	KNN Regressor	XGBoost
RMSE	161756	102003	156118	105344	102667	107301	`96308`	96361
R2	X	0.7512	0.3586	0.7320	0.7465	0.7108	`0.7636`	0.7571

1.7 Conclusion

- `Knn Regressor` and `XGBoost` ended up being our best models, but our results remain fairly poor with our predictions hovering around `$96,000`.
- It most likely won't be as accurate as consulting a local real estate valuation expert.

The current tool is best when utilized for separating houses into differents price brackets, facilitating the evaluator's task.

1.7.1 Possible reasons explaining those relatively poor performances :

- The american housing market is highly competitive and functions on a `bidding system` which means many houses are sold above their market value.
- May need additional Feature Engineering
- Very unlikely considering the dataset's quality but there could be faulty values
- Lack of meaningful variables

1.7.2 Next steps for project improvement

- Retrieval of additional variables (`price/m²`, parking_spots …)

- Testing `Polynomial Regression models`
- Prepare different models for different price ranges, and eventually run prediction from multiple models and make an average prediction of their results