HOUSE PRICE ANALYSIS

PROJECT

DANIEL SUAREZ SOUTO

dsuarezsouto96@gmail.com

Table of Contents

Section 1. Problem Statement

1. Business Problem

In this project we want to predict the prices of properties in the city of Washington. To do this, we have a dataset that contains information on 28900 properties in this city.

Some of the questions we want to solve with this project are:

- Which are the areas of Washington with the most expensive properties?
- Which is the tendency of the prices?
- Which are the features that have more impact in the prices of the properties?

Section 2. Data Preparation

1. Dataset

The dataset used is DC_PROPERTIES_TRIMMED, consisting of 48 columns and 28900 properties. The predictors have information of all kinds, from the date of sale, such as the year of construction, type of wall materials or neighborhood in which it is located**.**

2. Missing Values

Some of the categorical variables have missing values labeled "No Data". Others seem to have been manipulated by filling in the missing values with the label "Default". Each predictor that had these labels has been analyzed individually and these were the statistics.

Variable	Missing Samples
HEAT	5
GRADE	1

As we can see, the number of instances with some of the missing values was very small and so I decided to eliminate those instances.

3. Handle Errors

Once the missing values were resolved, I proceeded to discover errors in the data set. In the following list I summarize the errors found.

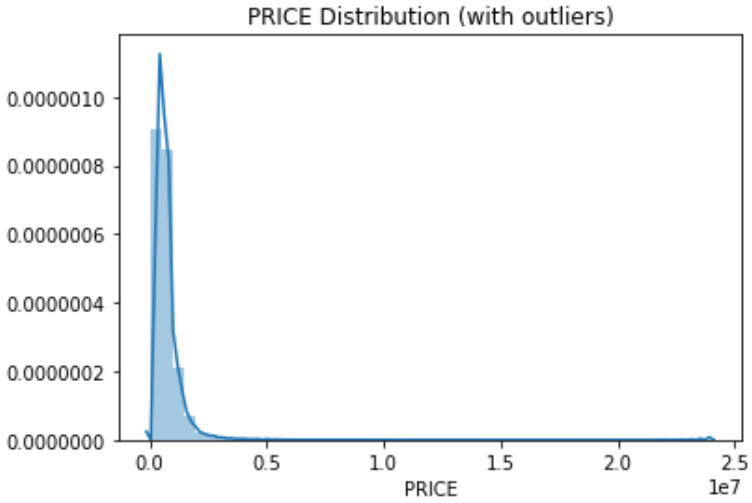
- AC: there are 8 instances with a value '0'.
- EYB: there is a sample where the EYB is before the AYB.
- YR_RMDL: there are 20 samples where YR_RMDL is before AYB, and one sample where the year of remodeling is 20.

Although these are not errors, I also want to include in this section that fictitious predictors like:

- STATUS: the same value for all instances (DC).
- CITY: same value for all instances (Washington).
- SOURCE: same value for all instances (Residential).
- GIS_LAST_MOD_DTTM: same value for all instances.
- X and Y: repeated with latitude and longitude.

4. Outliers

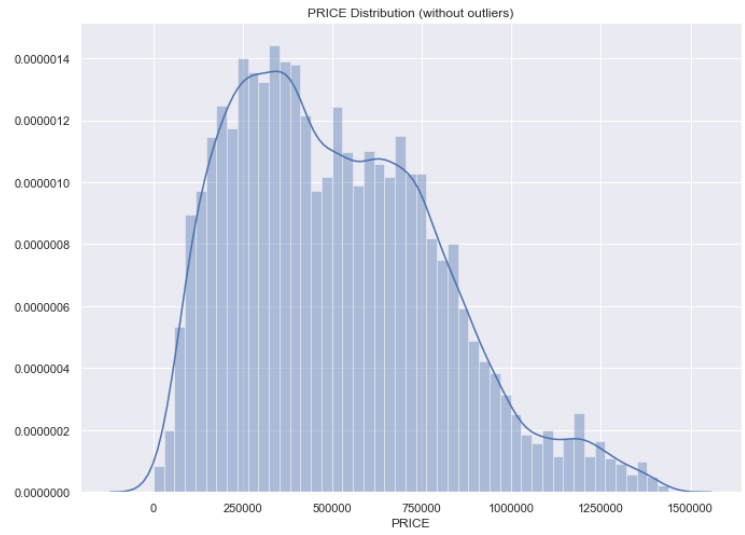
As mentioned in the previous section, my variable target is the PRICE of the properties. The following figure shows the price distribution of the properties.



As you can see, the distribution is totally influenced by the outliers of the property prices. So a process is needed to eliminate these outliers.

On the other hand, it does not seem reasonable to want to detect outliers in the price of properties, comparing prices without differentiating between their characteristics. The average price of a neighborhood with little purchasing power can be an outlier in the rich neighborhood. So my strategy to eliminate outliers has been to filter by neighborhood in which it is located and year of sale. These predictors have been chosen after a simple analysis among the variables, and which shows that these predictors are among the most discerning among property prices.

The number of outliers detected with this procedure has been 3884. The following figure shows the distribution of *PRICE* without the outliers.

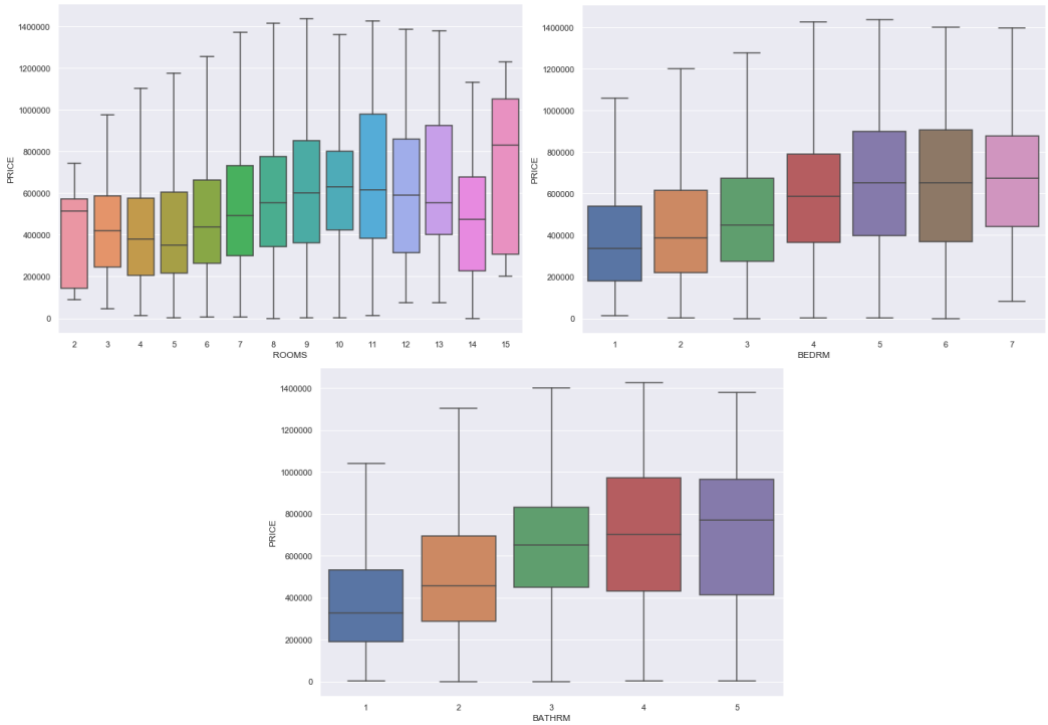


Section 3. Exploratory Data Analysis (EDA)

In this section I will apply the different analyses that I have done to study the relationship between the prices and the variable target. For a better explanation I have grouped some variables of similar meaning for their study and joint analysis.

5. Size Variables

In this group I have included variables related to the size of the household such as number of rooms or living area. The first analyses were done by analyzing the number of bedrooms and bathrooms. The following figures show their relationships:

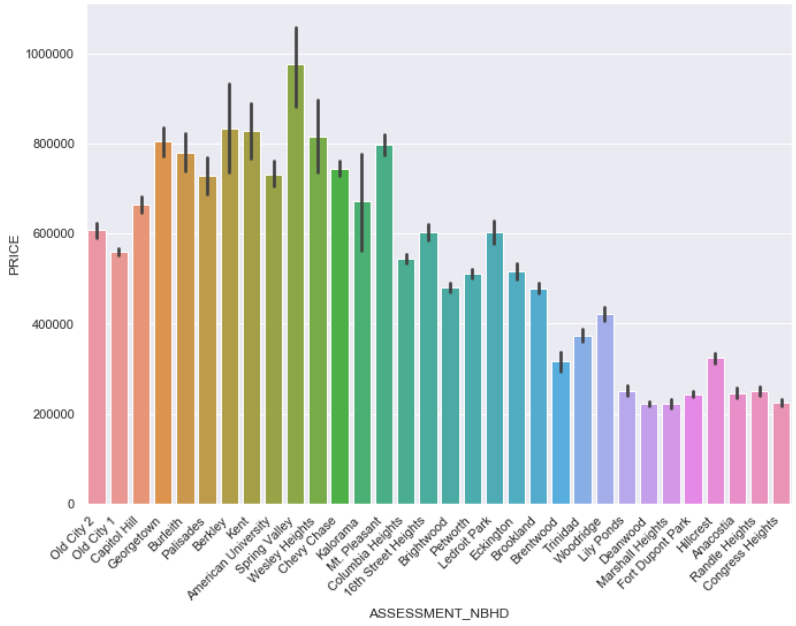


We can highlight a number of discoveries:

- The distribution of prices is practically the same in properties of 5 and 6 rooms.
- The increase from 1 to 2 bathrooms has the greatest impact on property prices (increase of 33%).
- The price of the properties depending on the number of rooms is different from the other two. The rooms with two rooms, although they have a more dispersed distribution have a fashionable value ($446500) 6% higher than with 3 rooms.

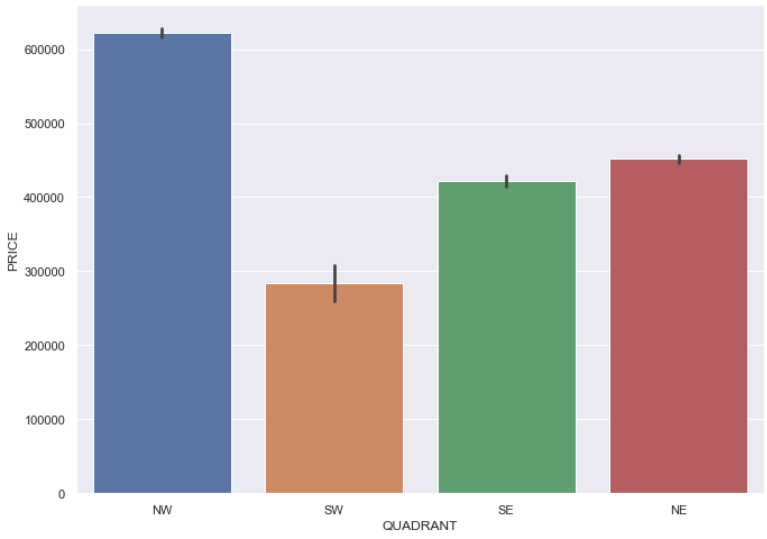
6. Location Variables

In this group of variables I have studied the impact that different areas have on property prices. As you can see, in the different neighborhoods there are high differences in the average prices of the properties.

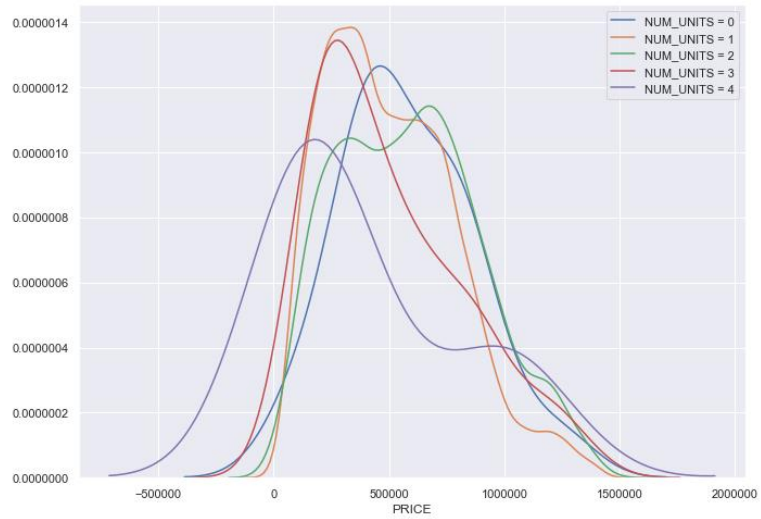


We see that the neighborhoods with the most expensive properties are Spring Valley and Palisades while the cheapest are Deanwood, Marshall Heights and Congress Heights.

The *QUADRANT* predictor allows us to analyze the territories with a lower level of precision, but it allows us to know if the neighborhoods with more expensive properties are closer to each other. We can see how the most expensive neighborhoods are in the NW quadrant of Washington. There is a considerable difference between NW and SW of more than $400000 on average between both quadrants.



Another hypothesis that I wanted to analyze is whether there is a trend for the price of properties according to the *NUM_UNITS* of the building. The following graph shows the distribution of *PRICE* according to the number of units.

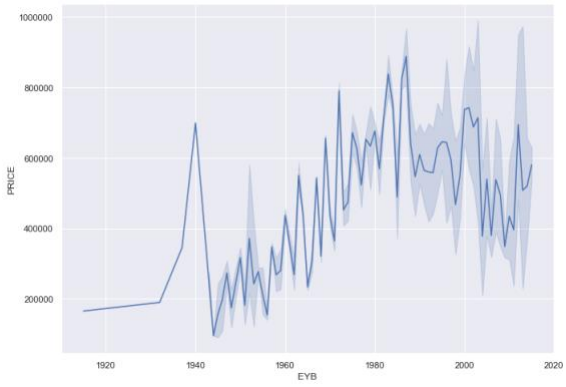


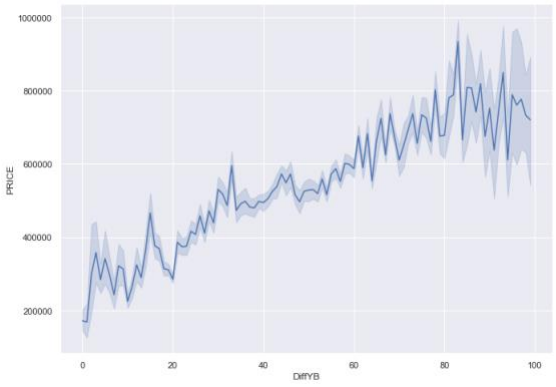
As we can see, the units with the lowest medians have the highest number of units (3 and 4). While the properties with number of units equal to 2 correspond to the ace that the highest. However, as can also be seen in the figure, the distributions are very overlapping and a considerable difference between them cannot be seen.

7. Building and Remodelation Year

In the dataset we have information on when the properties have been built or in what year they have had some remodeling. In addition, we also have the *EYB* information that corresponds to the year in which the last improvement in the properties has been made.

One of the predictors that I have created is *DIFF_YB* which corresponds to the difference between the year in which the last improvement was made *(EYB)* and the year in which construction started *(AYB).* We show in the following figures the relationship between *EYB*, *DIFF_YB* and *PRICE*.

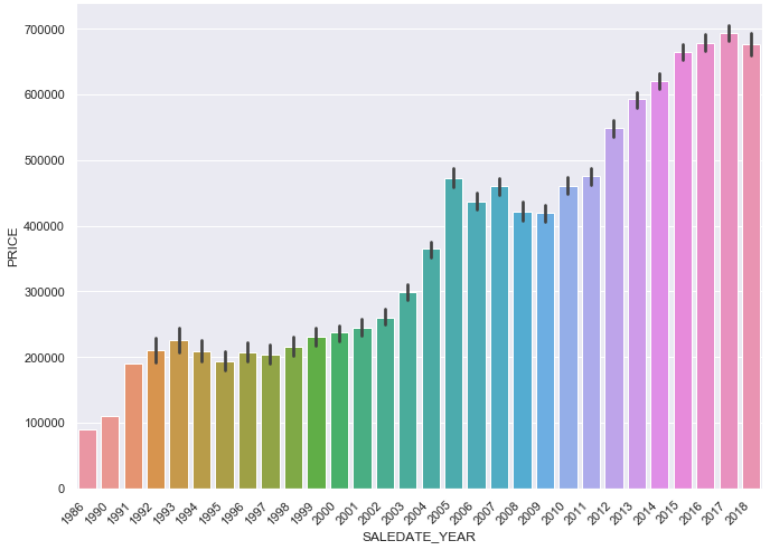




We can see that there is a clear correlation between the variable created and the price of the properties. It should be noted that we have limited the representation of *DIFFYB* to 100 years. It is remarkable that there are properties where the difference between EYB and AYB is so big. It may be due to the diffuse difference between "improvement" (which corresponds to EYB year) and "remodeling" (which corresponds to YR_RMDL) and it may lead to some remodeling being stored in EYB.

8. Sale Date Variable

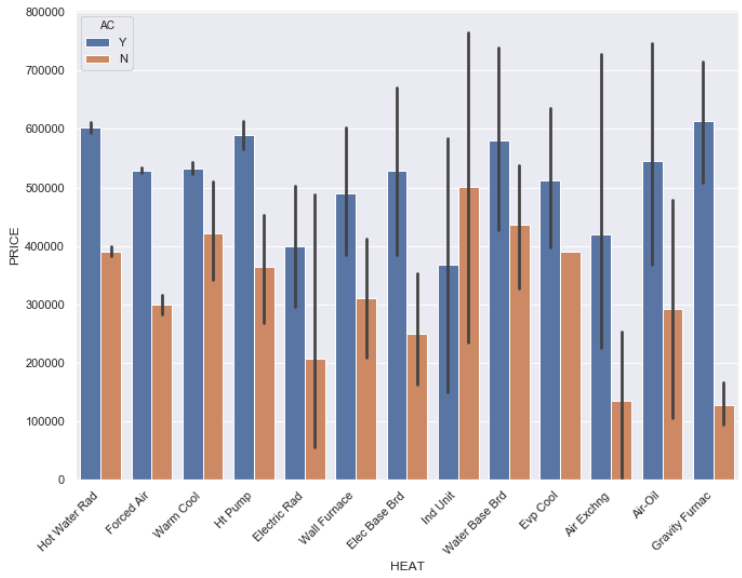
The date of sale seems to be an important factor in determining the price of properties. The real estate market is still an investment market, where the value of the properties depends considerably on the context in which they are located. The following chart shows the price per household according to the year of sale.



It is worth noting that the average has always been positive, with the most recent years being those with the highest average property prices.

9. Heat and AC

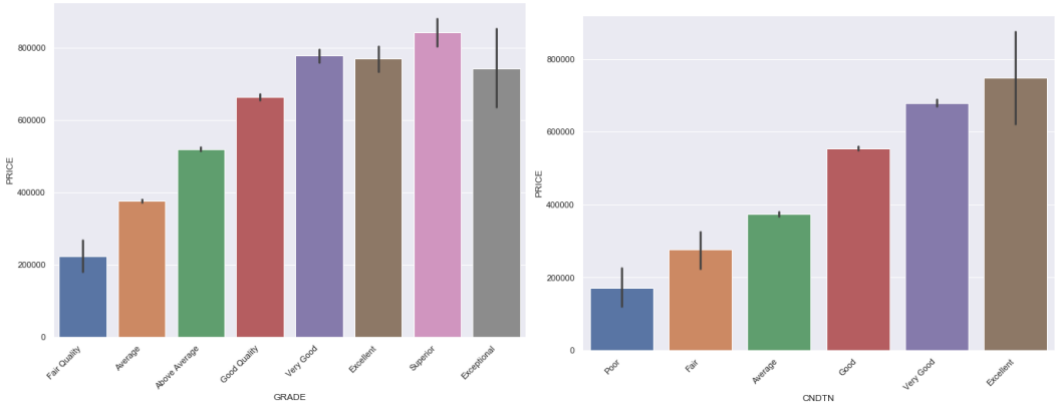
These variables determine whether or not the properties have air conditioning and the type of heating that is counted.



The first thing we notice is that properties with air conditioning are more expensive than those without it. The heating types of the most expensive properties are those with a Hot Water Radiator and Gravity Furnace system.

10. Grade and Condition

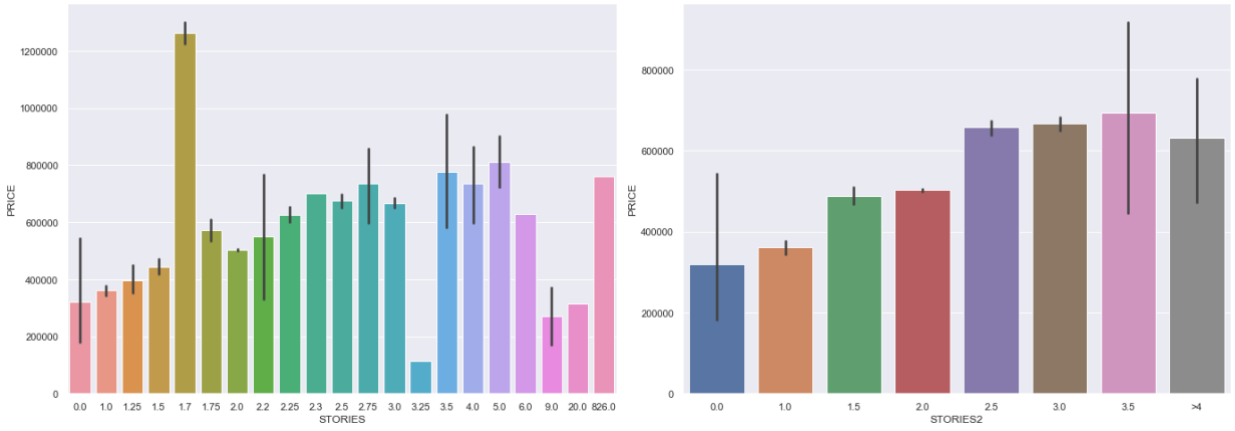
These *Ordinal Variables* contain information about the conditions in which the property is located. One of the modifications I have made is to join the different levels of "Exceptional - X" of the predictor GRADE in a single value corresponding to "Exceptional".



As we can see there is a clear relationship between the state of the properties and the price.

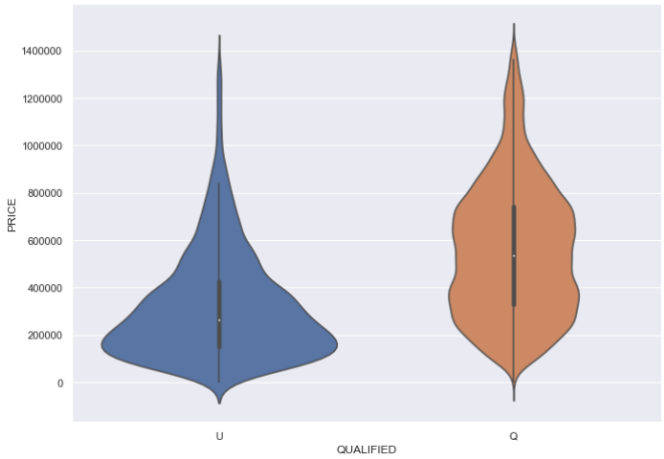
11. Stories

Story is any level part of a building with a floor that could be used by people (for living, work, storage, recreation, et cetera). The distribution of this variable is shown in the following figure. As can be seen, its values were scattered (with values such as X.3). For that reason I decided to make a new categorical variable that had the traditional values, and as it can be appreciated the relation seems to have increased.



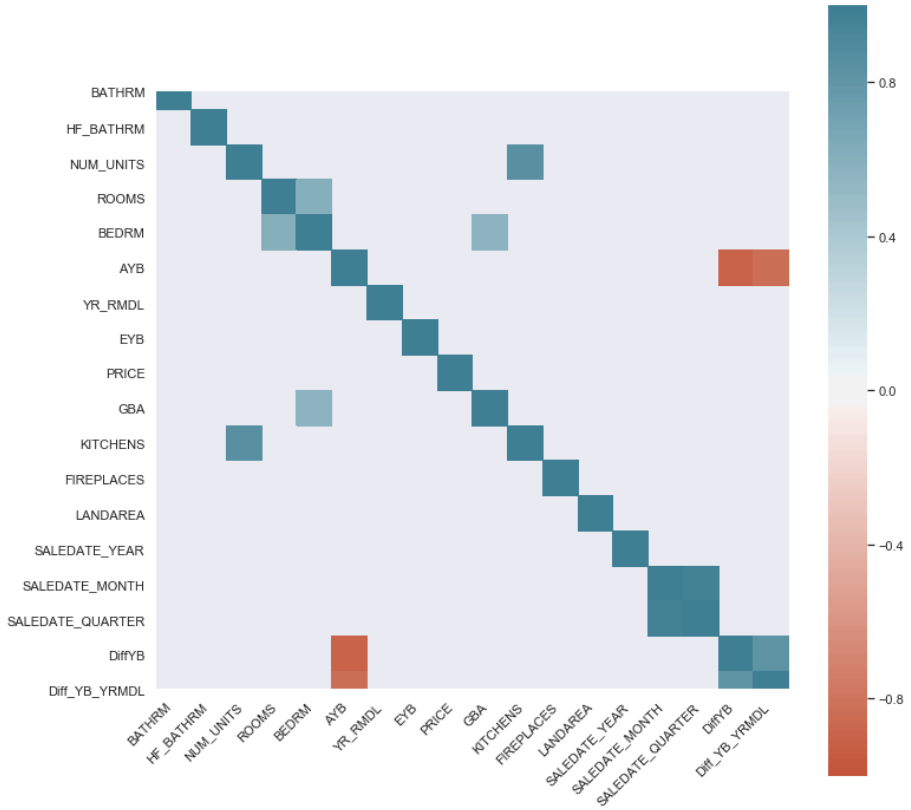
12. Qualified

This variable defines whether or not the properties are qualified according to certain parameters. As you can see in the following figure, the price distribution for qualified properties contains larger values than those that are not qualified.



13. Multicollinearity

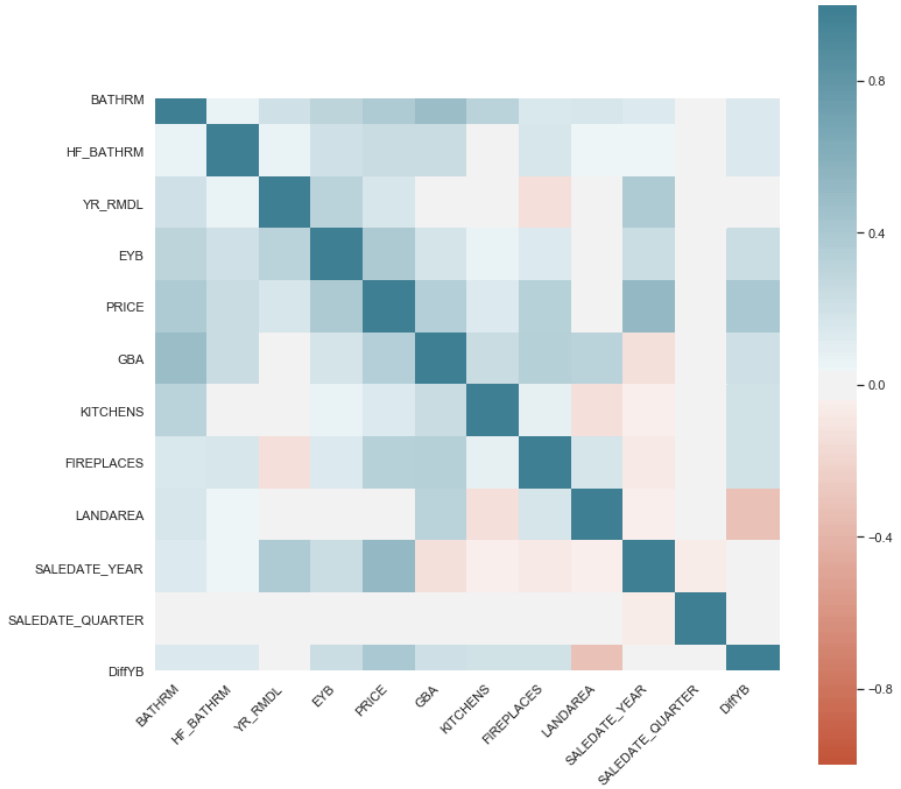
The following figure shows the correlation of numerical variables that have a high value (greater than 0.55 in absolute value).



As we can see, there are variables that have a high correlation between them, so to say which variables to use, a statistical study of them has been carried out, summarized in the following table.

Variables		P-Value	Correlation with Target Variable	Variable Selected
NUM_UNITS vs KITCHEN	NUM_UNITS	3.23e-39	-0.014	**KITCHEN**
	KITCHEN	3.23e-97	0.018	
ROOMS vs BEDRMS vs GBA	ROOMS	7e-277	0.23	**GBA**
	BEDRMS	0	0.25	
	GBA	0	0.35	
AYB vs DiffYB vs Diff_YB_YRMDL	AYB	5.18e-279	-0.23	**DiffYB**
	DiffYB	0	0.4	
	Diff_YB_YRMDL	0	0.29	

After selecting one from each group of variables this is the correlation matrix.



Section 3. Modelling

For the prediction of the price of the properties different models are going to be tested. The models with which experiments have been carried out are:

- **Random Forest**: Bagging algorithm that ensembles a number of decision trees.
- **Ridge**: Linear regression algorithm with L2 regularization.
- **Ridge Transformed**: as it could be appreciated, the distribution of the target variable seems to be skewed. This algorithm applies a log transformation in the target variable and use a Ridge estimator.

14. Preprocessing

Although the dataset is clean and with the variables that we want to use for the estimation of the price of the properties, we have made the following processes for the pre-processing of the features:

- **Coding Variables**: for the Categorical Variables we have made One Hot Encoding to code them, while for the ordinal variables we have used Potty Encoder, to keep the order between them.
- **Standardization**: because we use the Ridge algorithm, the features are standardized before being used by this algorithm so that all are at the same scale.

15. Methodology and Results

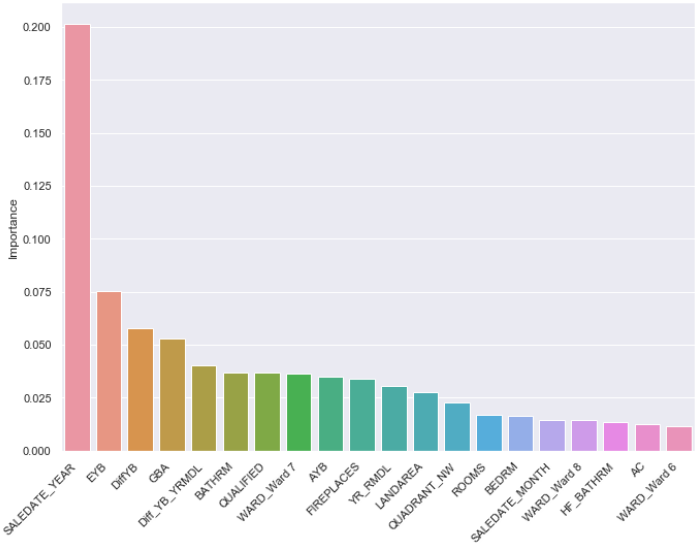
For the development and evaluation of the models a train-test split of the dataset has been made, being a 30% the test set. For the validation of the model, I have made a Cross Validation in the train set.

The following are the results of the three models in the train set, using KFold with k equal to 5.

Model	Train Set	
	MAE	R Squared
Random Forest	88286.73	**0.82**
Ridge	98428.87	**0.8**
Ridge Transformed	87092.22	**0.82**

16. Feature Importance

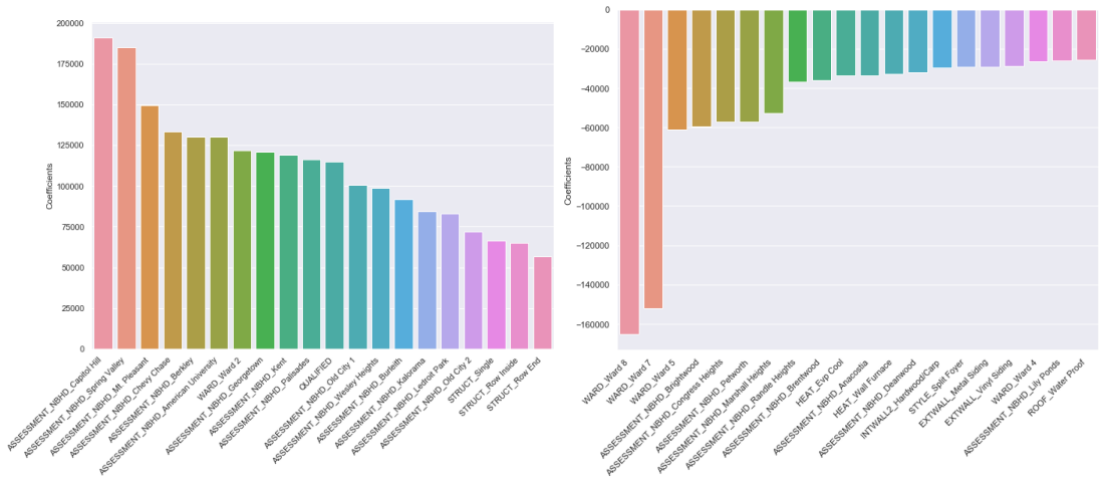
Random Forest algorithm allows us to know the importance of each predictor when estimating the price of a property.



It is worth noting the importance of variables such as *SALEDATE_YEAR* or *DiffYB*, which have a very high level of importance and have been created in the feature engineer process. In addition, as already mentioned in the analysis, *BATHRM* has a higher importance than the rest of the rooms in the house.

17. Coefficients

The following graph shows the 20 highest Ridge coefficients (both positive and negative).



The importance of the different neighbourhoods in determining the price of a house is obvious. Many of the coefficients with the highest value correspond to different neighborhoods.

18. Optimization and Performance

The GridSearchCV method has been used to optimize the hyperparameters of the algorithms and the models have been tested in the test set. The results are shown in the following table.

Model	Hyperparameters	Test Set		
		MAE	RMSE	R Squared
Random Forest	N_estimators : 1000	87518.88	119781.05	**0.826**
Ridge Transformed	Alpha : 10	98791.37	128505.74	0.80