HPEM

House Price Estimation Model

Cleaning Of The Dataframe

Case Study -We are an Data analysts working for a real estate company

Our job is to develop a model that predicts the price of houses to explore the characteristics of the houses over 650K

.Shape

First of all

I used the shape function (`raw.shape`) to show me the shape of the dataset

.Info() Function

We used the info() method that prints information about the Data Frame. The information contains the number of columns, column labels, column data types, memory usage, range index, and the number of cells in each column (non-null values).

.Describe()

The describe() method returns description of the data in the Data Frame.

It basically shows the rows and columns of a dataset

Checking for Nan`s , Checking Duplicate values ,Dealing and Rechecking

Isna().sum()

this function returns the number of missing values in each column and that way i can deal with Nan Values

We used **drop_duplicates(['id'])** to drop and deal with duplicate values then i used **['id'].duplicated().any()** to recheck for any duplicate value to be sure it was removed

Outliers

We used the Quantile ranges method to be able to detect outliers

We've identified a total of 30,6% of data that is considered outliers. This is almost a third of the dataset, so it wouldn't be smart to drop them all.

- Outliers may be important in identifying trends or patterns in the data that would be missed if they were removed.
- Also, the most expensive houses, >650k, may be outliers since they are a minority and have more features than the average houses.
- Since this data is based on genuine observations, we've decided to keep the outliers for now.

EDA

- The graphs represent the influence on price by the variables being analyzed.
- A **warmer tone** represents more **concentrated data** around that area. This means that possible **outliers must be blue**.

Few Findings

Once a house has a waterfront view the price

is really high

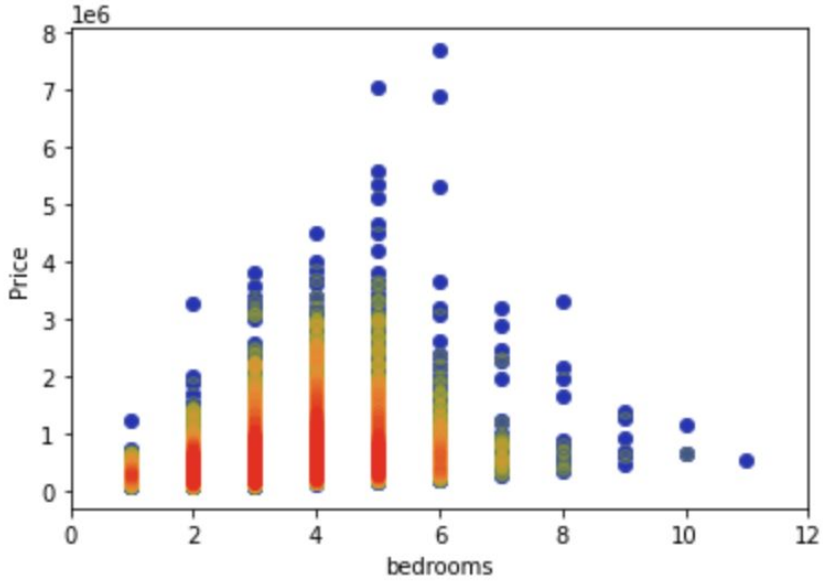
The most expensive house are in the medina Area

The higher the grade of the house higher the price

Influence of Discrete Variables on Price

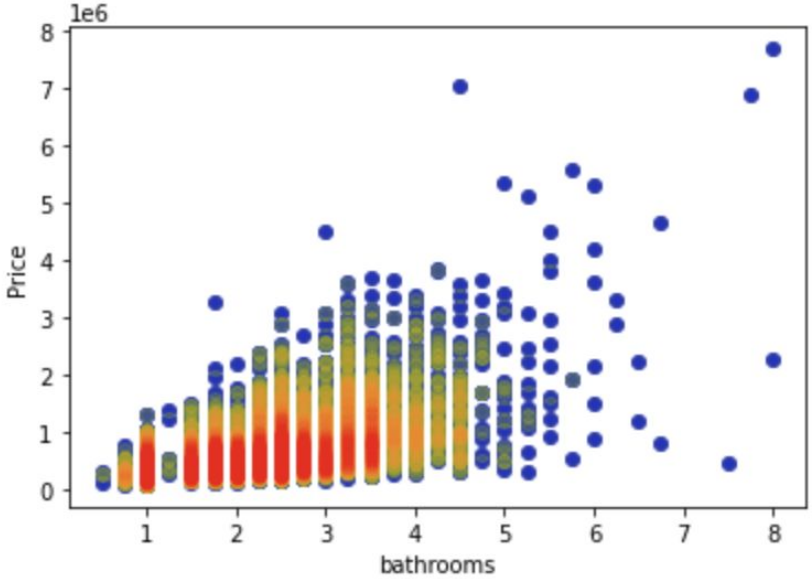
Bedrooms

- X axis limited from 0 to 12;
- Ratio price-bedrooms picks on 5 bedrooms;
- Data is concentrated around 2-6 bedrooms.



Bathrooms

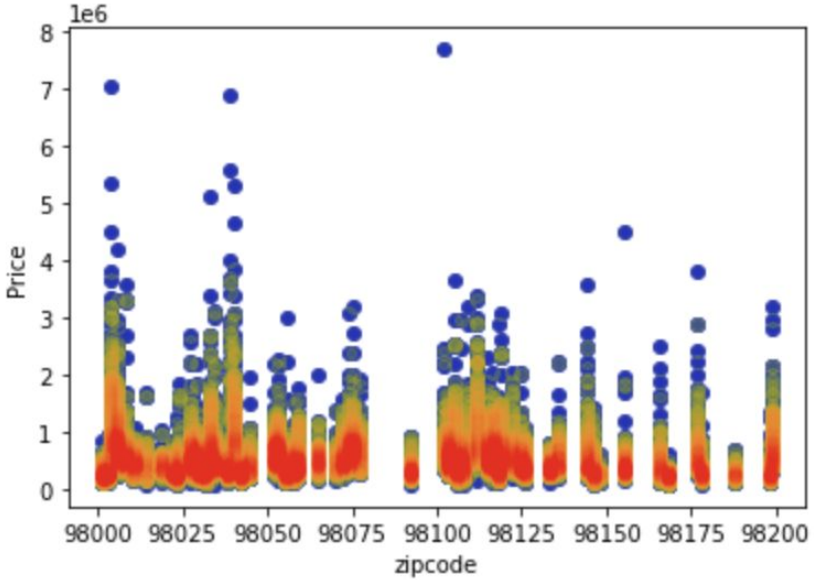
- Positive correlation;
- Exists a tendency to have a higher price for a bigger number of bathrooms;
- Data is concentrated on 1-4 bathrooms.



IRON HACK

ZipCode

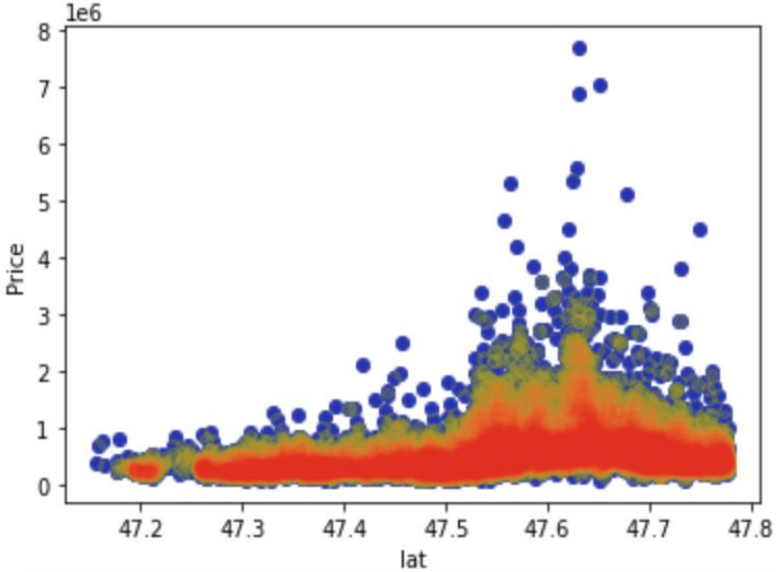
- Exists a tendency to have a higher price around some zipcodes.

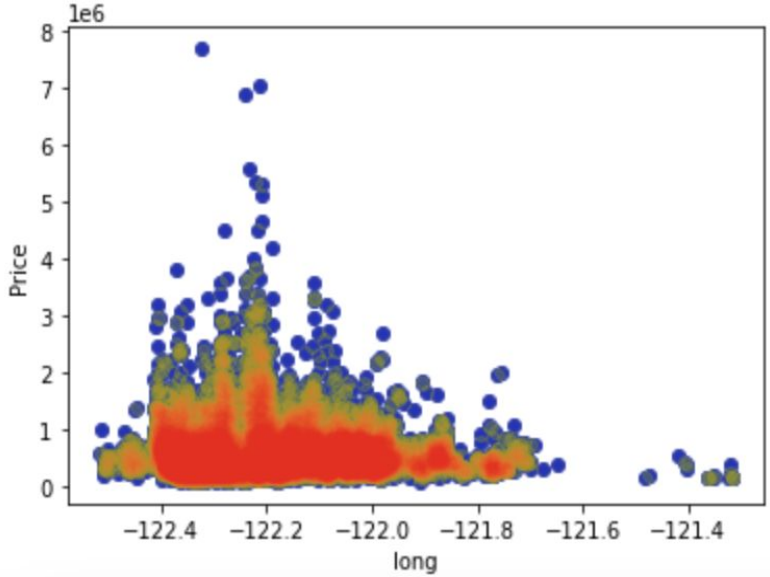


Influence of Continuous Variables on Price

Latitude and Longitude

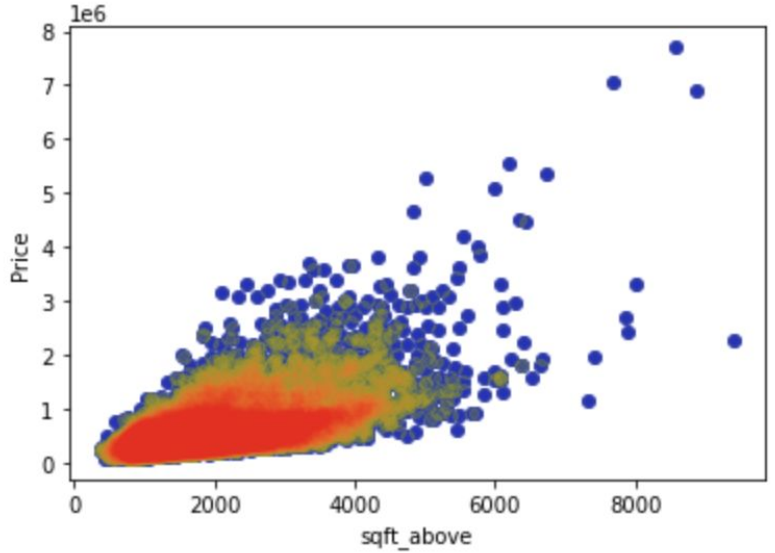
- These variables are connected;
- They change with the zipcode.





Highly Correlated Variables with Price

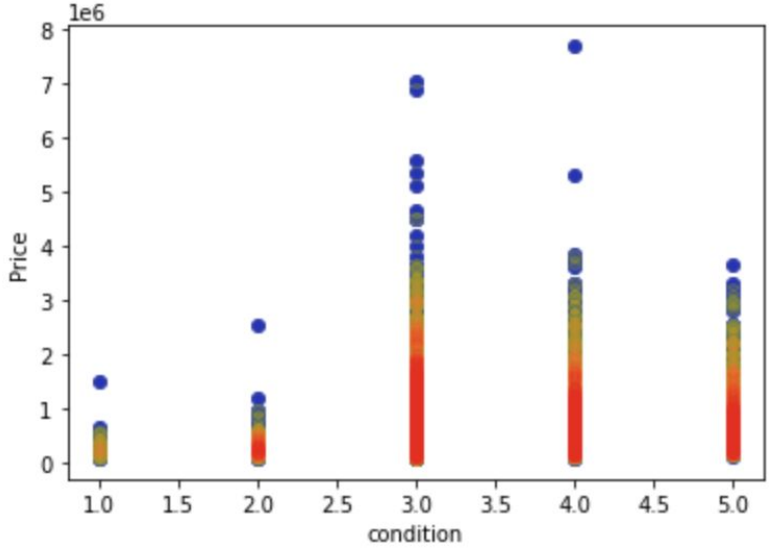
- 'sqft_living',
- 'sqft_above',
- 'sqft_basement',
- 'sqft_living15'.

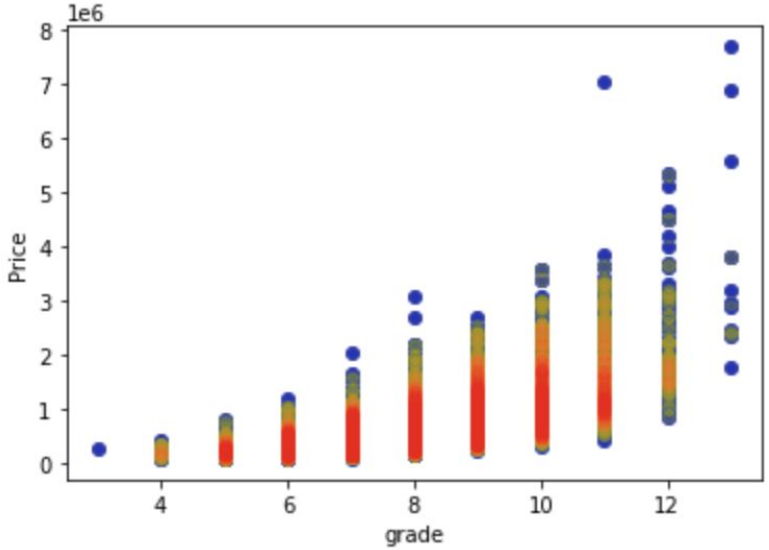


Influence of Categorical Variables on Price

Correlated Variables with Price

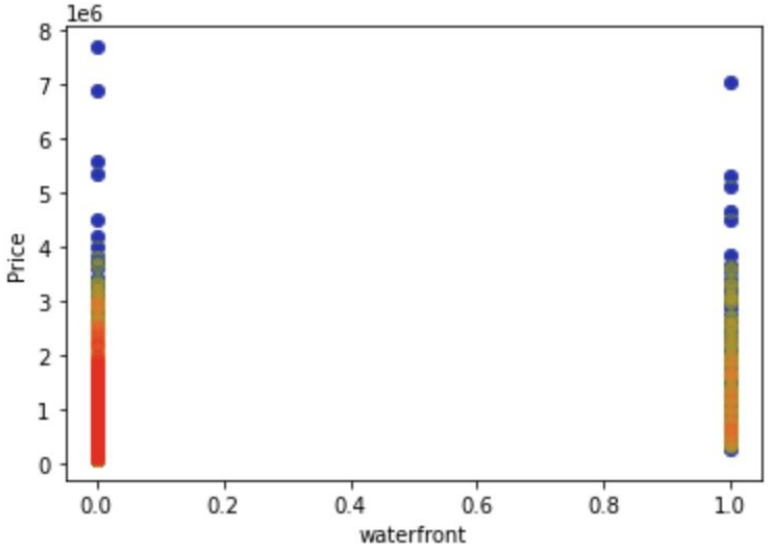
- ‘condition’;
- ‘grade’.

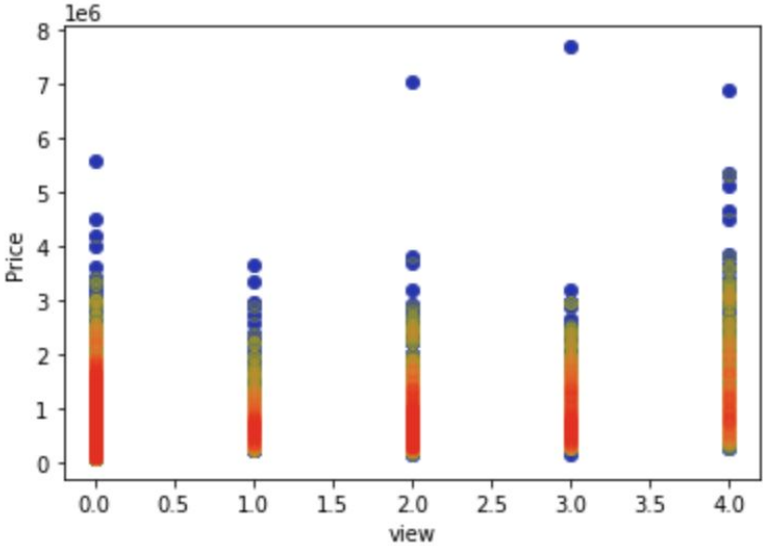




Non Correlated Variables with Price

- ‘waterfront’;
- ‘view’.

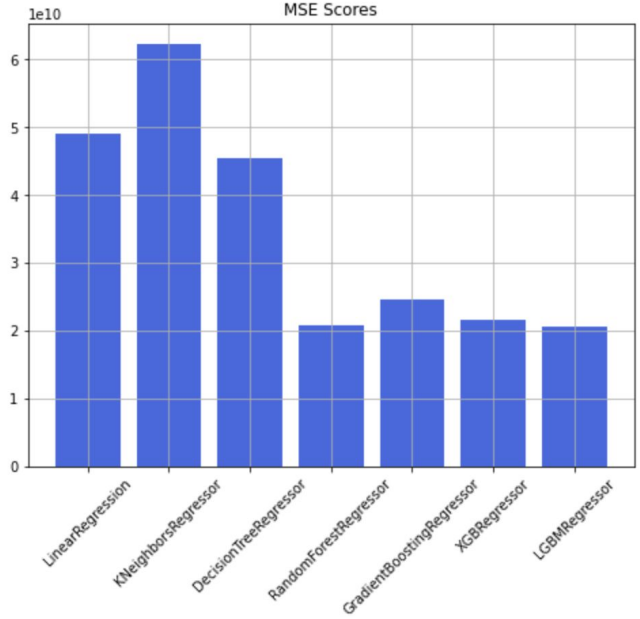


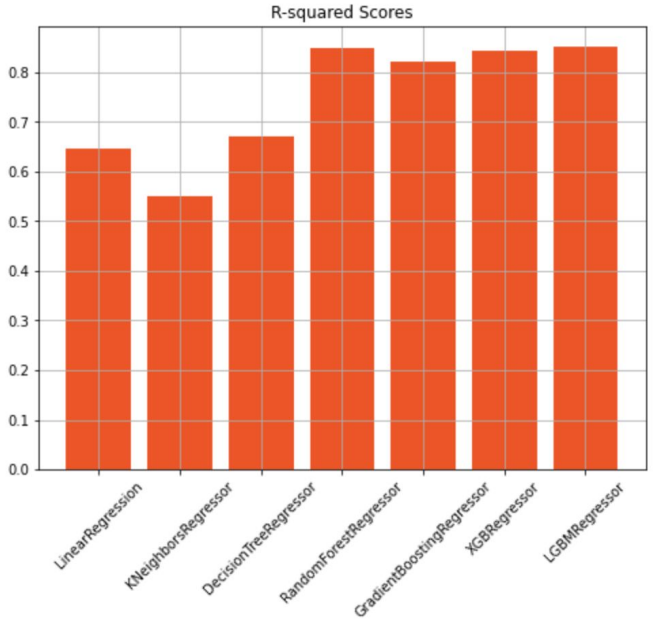


Modeling

- LinearRegression;
- KNeighborsRegressor;
- DecisionTreeRegressor;
- RandomForestRegressor;
- GradientBoostingRegressor;
- XGBRegressor;
- LGBMRegressor.

Comparing Models Performance





Conclusions

- Random Forest Regressor and LGBM Regressor did the best predictions.

Questions?

Thank You!