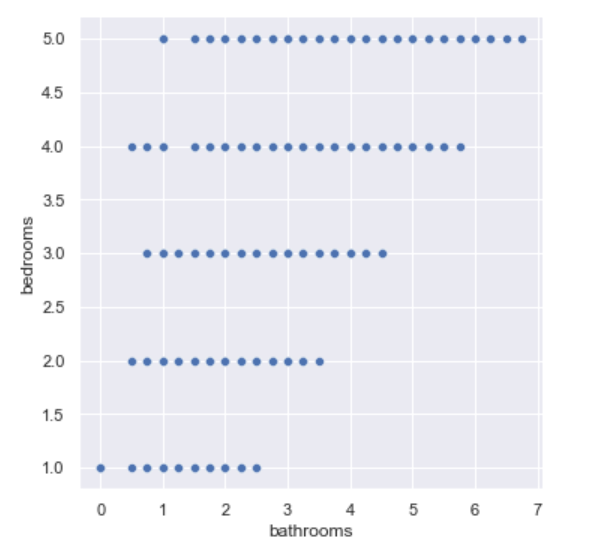
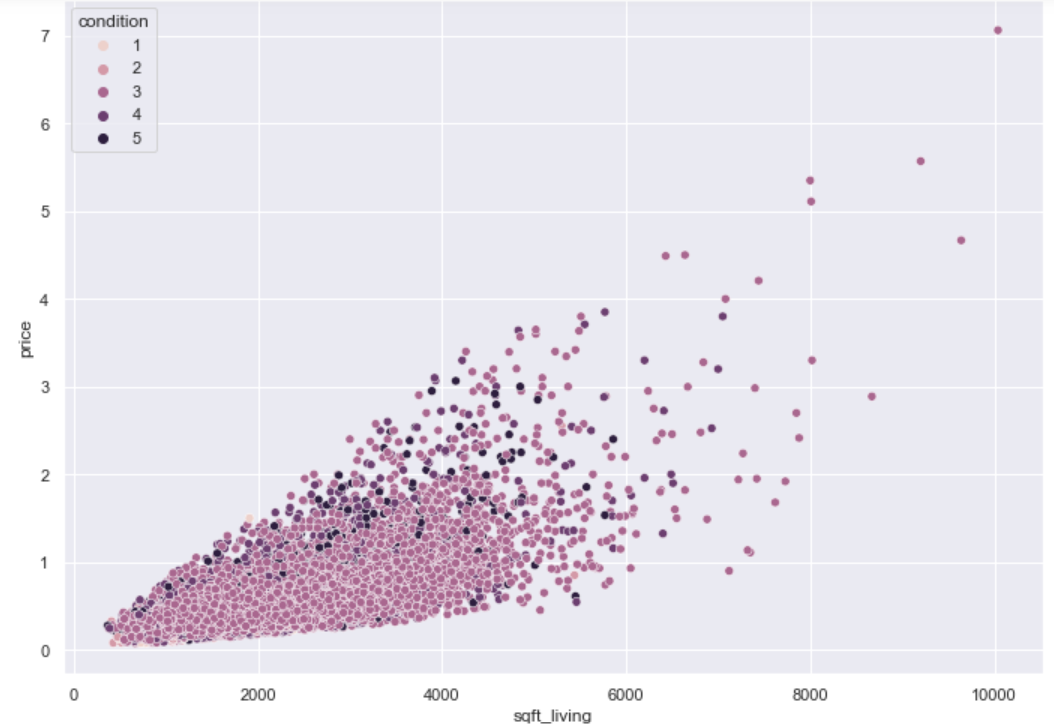
**ML ASSIGNMENT**

* Larger the houses, larger is the number of bathrooms, ie five and four bedroom houses have the most number of bathrooms.

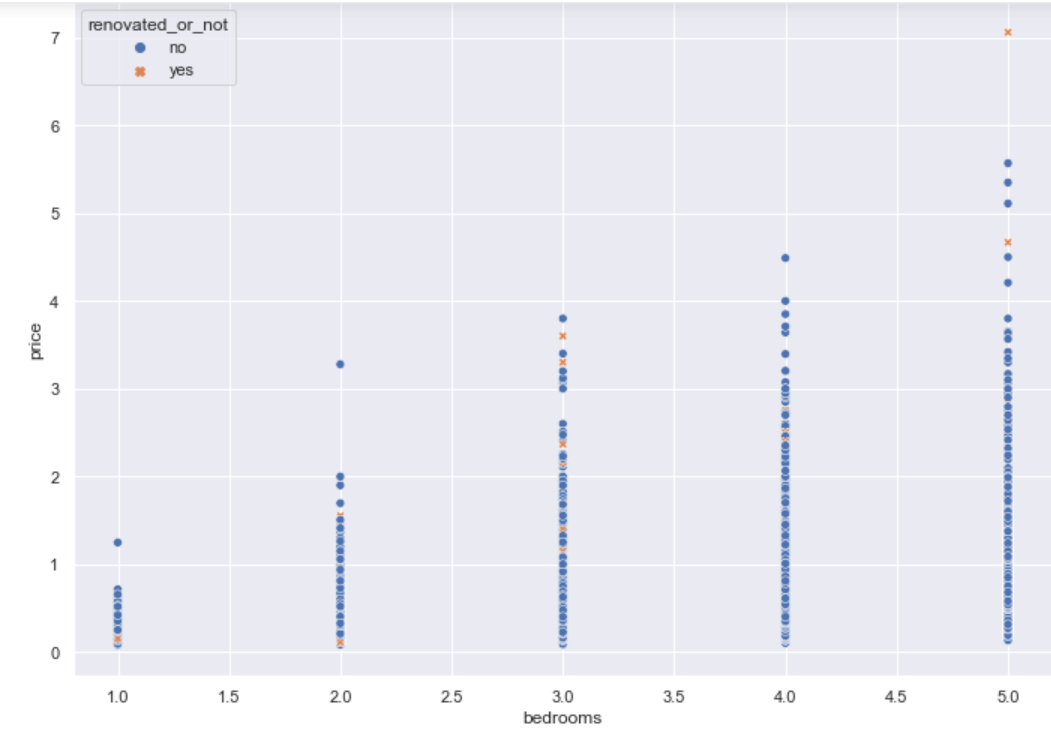
****

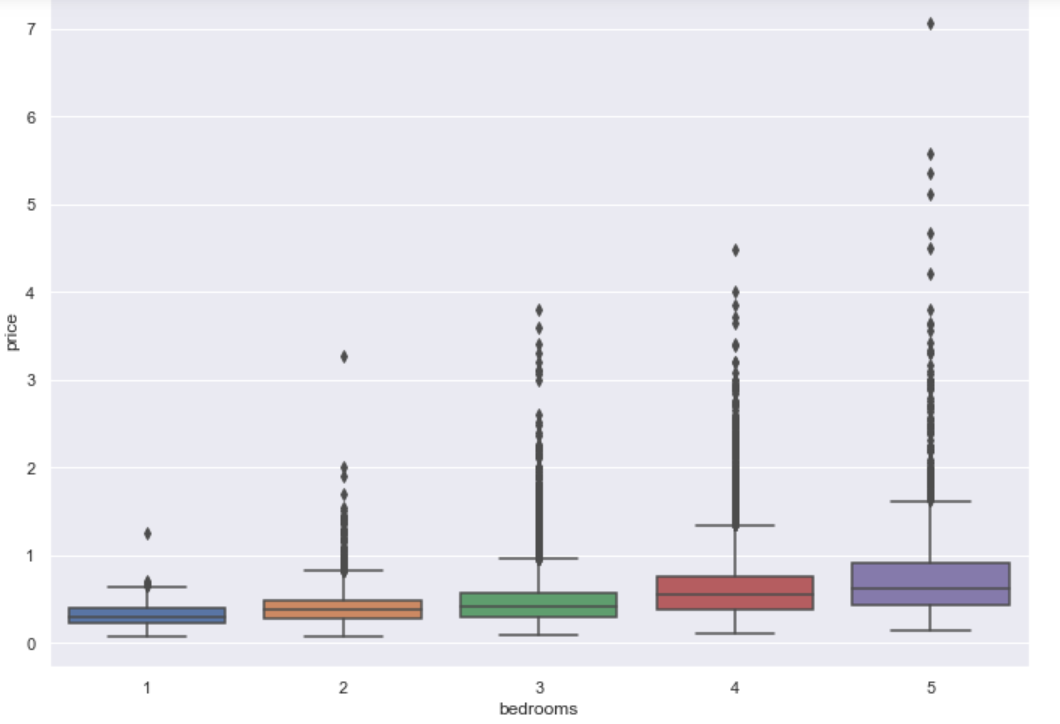
* There are almost no houses with a condition of 4 or 5 and living area greater than 8000 sqft, ie most of the houses with a good condition have a sqft area of living not more than 8000 also there are very less houses with the maximum condition 5.

The highest priced houses are not seen to have a condition of 5.

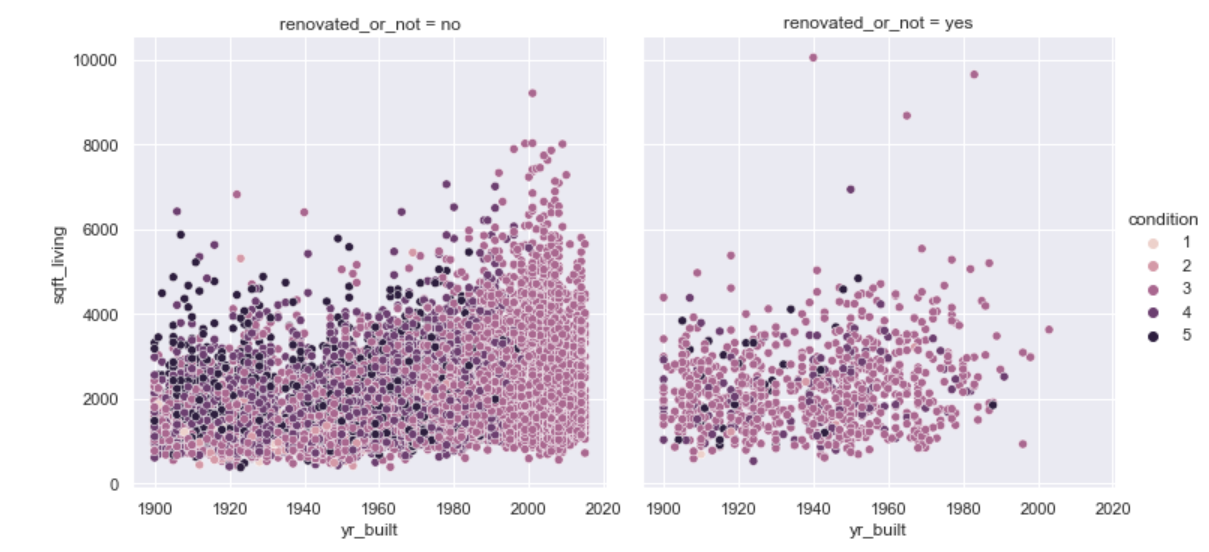


* Highly priced houses are 5 bedroom houses, there are very few houses that have been renovated, most of the houses haven’t been renovated.





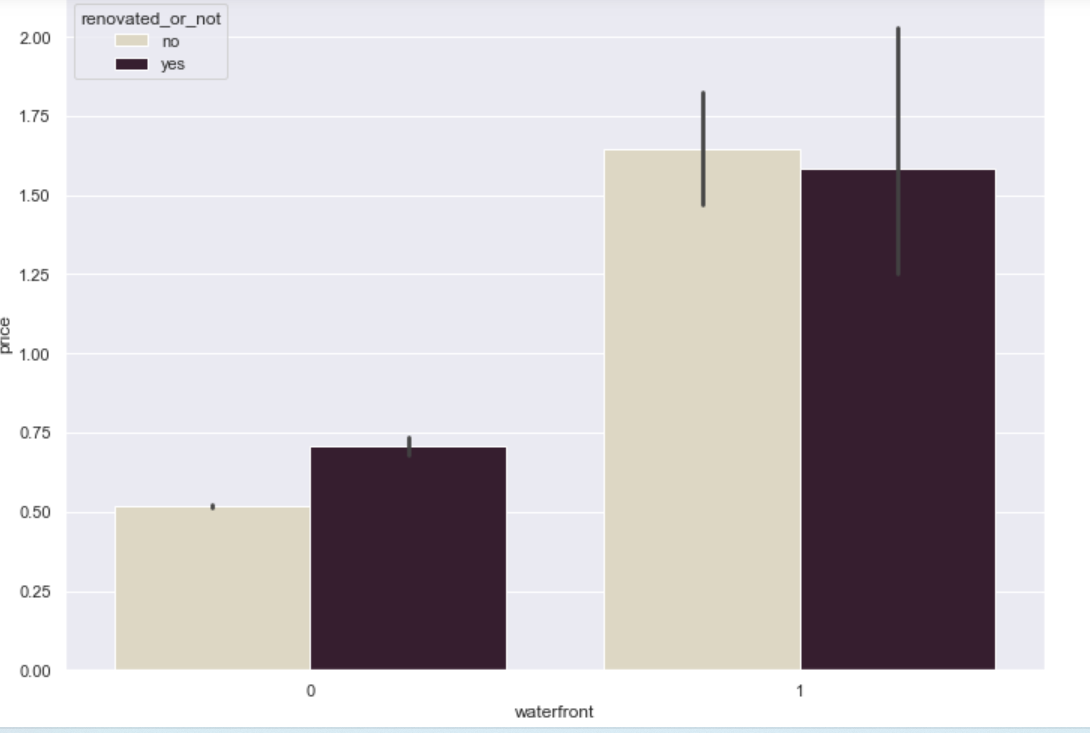
* Most of the houses are not renovated, and the ones which are renovated are mostly built between 1900 and 2000, also better conditioned houses are the ones which are not renovated.



* Houses with a condition one or two have square feet of living in the range of 1000 to 4000, there are very less houses which have been renovated and most of the houses have a condition of three(most houses are average conditioned).



* Houses having a waterfront are highly priced. Majority of the houses without waterfront have been renovated and the ones with waterfront have equal proportion of renovated and not renovated ones with a slightly high proportion of renovated houses.



* What are the assumptions of linear regression?

Linearity: The relationship between X and the mean of Y is linear.

Homoscedasticity: The variance of residual is the same for any value of X.

Independence: Observations are independent of each other.

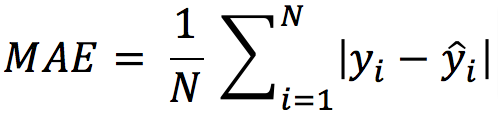
Normality: For any fixed value of X, Y is normally distributed.

* How can we evaluate a Regression model? Define each metric and its interpretation.

Regression predicts a continuous dependent element in the presence of various independent elements. Linear regression tries to make a trend line that has the least difference between actual and predicted values. This difference is also known as residual.

**MEAN ABSOLUTE ERROR (MAE)**

It is the mean of the absolute difference between the actual value in the dataset and the value predicted by the model.

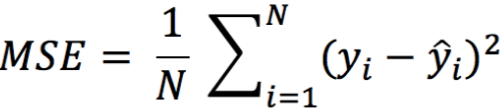


N = the count of the data points.  
 y = the actual value in the dataset  
 y cap = the model’s predicted value.

The absolute values are taken, and if it’s not then the negative and positive difference will cancel out each other. The smaller the MAE, the more accurate the model is. If MAE is zero it shows the model is perfect. If MAE is large then the model is not good.

**MEAN SQUARED ERROR (MSE):**

This is the mean of the squared difference of the actual value in the dataset and the value predicted by the model.



N = the count of data points in the data.

y = the actual value in the dataset.

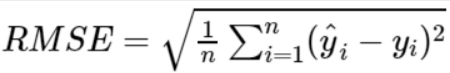
y cap=  the model’s predicted value.

The MSE will be large if there are outliers in the dataset, this is not the case with MAE.

MSE focuses on larger errors, as when we are squaring the error the effect of large errors becomes more prominent. If the errors are low, lower than one, then it leads to underestimating the model’s error.

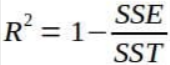
**ROOT MEAN SQUARED ERROR:**

It is the mean of root squared subtraction between the actual value in the dataset and the value predicted by the model.It’s the same as MSE, we are just taking the root of it.The smaller the value of root mean squared error, the more accurate the model is.



**R SQUARE**

It estimates the ratio of the variance of the dependent element described by the target element. It’s used for finding the accuracy of the model. It depicts the closeness of the data points to the trend line made by the model. This helps to make a link between the independent element and the target element. R square is from zero to one. The nearer R square is to one, the more accurate the model.



SSE is the sum of the square difference of the residuals.

SST is the sum of the difference of the actual value of the data and the mean of all the actual values in data.

* Can R squared be negative?

Yes, R squared can have a negative value

* What is dummy variable trap?

When the number of dummy variables created is equal to the number of values the categorical value can take on. This leads to multicollinearity, which causes incorrect calculations of regression coefficients and p-values.

* Is One Hot Encoding different from Dummy Variables?

[One-hot encoding](http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html) converts it into n variables, while [dummy encoding](https://en.wikiversity.org/wiki/Dummy_variable_(statistics)) converts it into n-1 variables. If we have k categorical variables, each of which has n values. One hot encoding ends up with kn variables, while dummy encoding ends up with kn-k variables.

* How is polynomial regression different from linear regression?

Simple Linear Regression establishes the relationship between two variables using a straight line. It attempts to draw a line that comes closest to the data by finding the slope and intercept which define the line and minimize regression errors. Simple linear regression has only one x and one y variable.

Polynomial Regressionis a one of the types of linear regression in which the relationship between the independent variable x and dependent variable y is modeled as an *nth*degree polynomial. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted E (y |x).

Polynomial Regression provides the best approximation of the relationship between the dependent and independent variable.

* Interpret the screenshot below from the notebook we discussed in class today

Return the coefficient of determination R^2 of the prediction. The first code talks about the accuracy between the actual test data of the independent variable and the predicted values of the test data of independent varable. 1 is the maximum value,it indicates perfect fit.

The second piece of code also indicates prefect fit between the actual training set of data of the independent variable and the predicted value of the independent variable.

The third value is not one but nearly one and it indicates high accuracy but not perfect fit between the test data of the dependent variables and response variable.

The fourth value also indicates high accuracy but not perfect fit between the train data set of the predictor and response variables.

SUBMITTED BY:

PRAKRUTHI B GOPAL

21BDA25