Individual Assignment 2

Name: Tran Hy Dong

SID: 500286001

1. INTRODUCTION

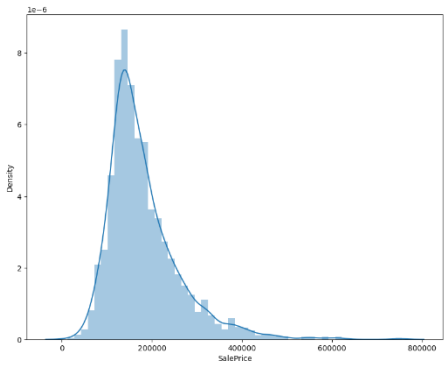
This report analyses the linear multiple correlation that exists between house prices in Ames, a city in state of Iowa and multiple factors related to both internal and external factors which can be considered as a determinant to affect house price in that area. Therefore, setting up a sales forecasting model to predict the Sales price to make a good prediction of each house which avoid the future asymmetric information in evaluate house prices in real estate industry which hugely contributed for the economic recession in 2008. To make the prediction, 3 models of forecasting is applied which is Multiple Linear Regression with different factors. After all the calculations and the comparison between 3 models, the combination of both internal and external factors should be more taken into account in predicting house price because it fit well with the dataset due to its high adjusted R-squared and the forecast accuracy measured by RMSE & MAD.

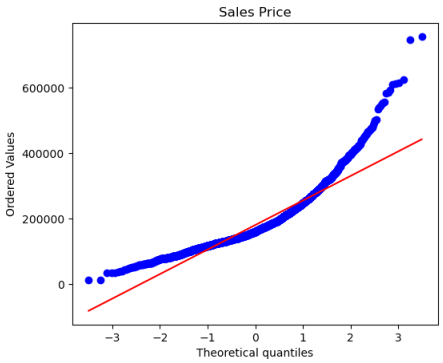
The Ames Housing dataset describe the sale of individual residential property in Ames, Iowa from 2006 to 2010. The data has 82 columns which include 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables (and 2 additional observation identifiers) from 2930 entries for each

2. CANDIDATE MODELS

To choose the determinants across all the, we must take a look of all the correlation between the Sales price with all parameters in the data set:

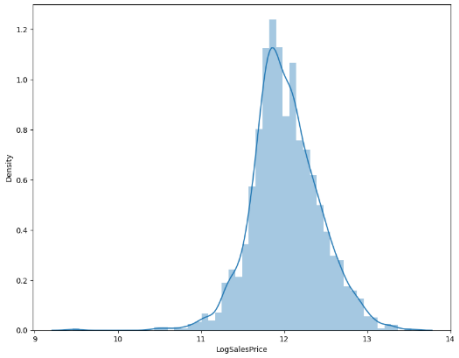
a. Our predicting variables: Sales Price

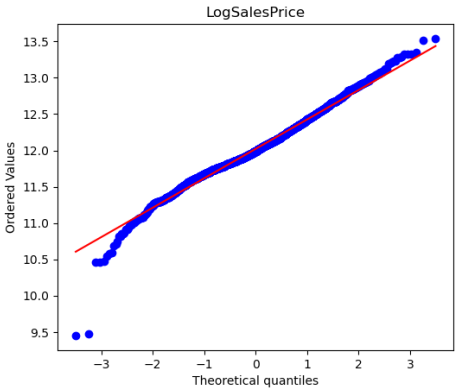




This shows that Sale Price does not follow normal distribution and has a long-tail distribution. It has positive skewness. It means that most of the house are normally distributed but a couple of houses have a higher-than-normal value of the Sale Price is not linear, which means we cannot find of a straight line that would fit through. Hence, we will try to transform to log sales price. To compare between the Sales Price and Log Sales Price, the distribution of Sales Price is right

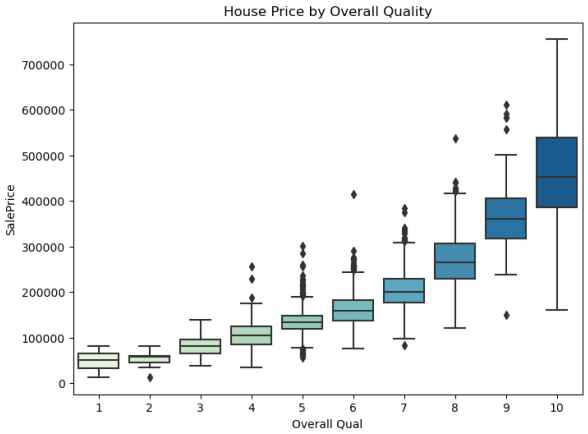
skewed, but after log-transforming, the skewed value is -0.014, which close to 0 stand for normal distribution. If we want to do the price-predicting model, log Sales Price should be the variables to be chosen to be predicted.

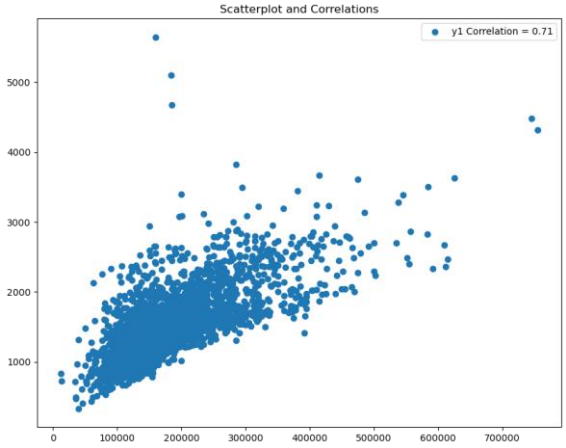




We can see that the points lie mostly along the straight diagonal line with some minor deviations along each of the tails. Based on this plot, we could safely assume that this log Sale Price is normally distributed. So, we will use the Log-Sales Price for predicting our Sales Prices.

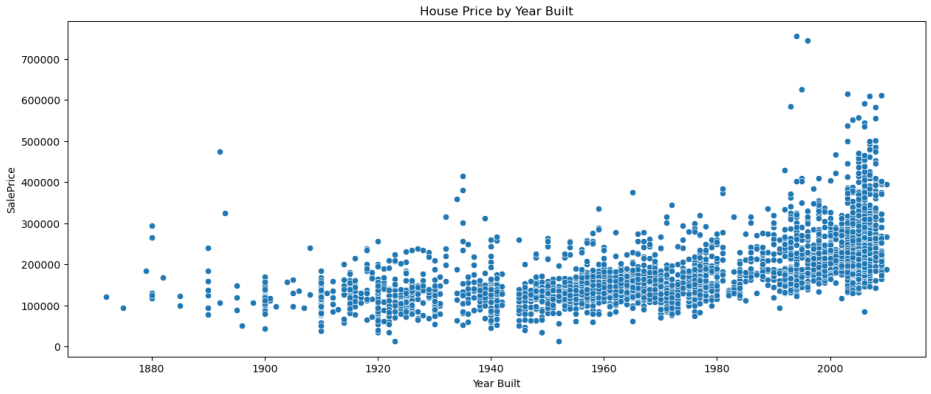
b. High Numerical correlation determinants



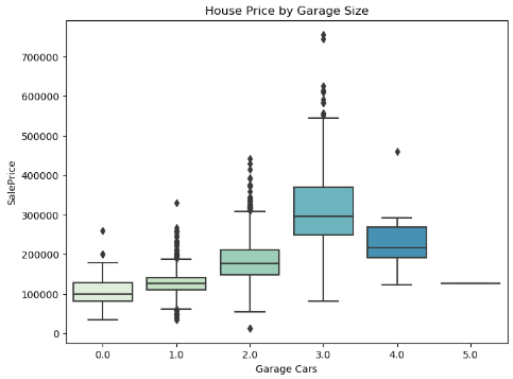


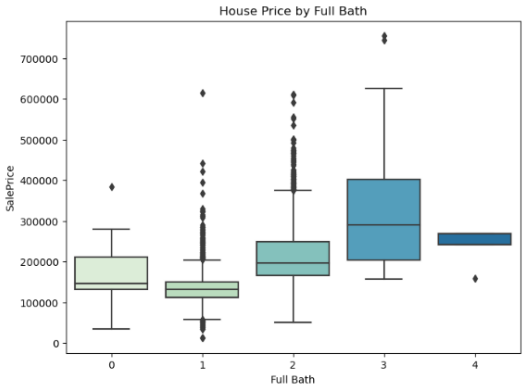
Overall quality is the most important feature in both analyses with high correlation stated at 0.8. It is clear that higher quality makes the house more expensive. Following is Living area has a linear relationship with house price. In the scatter plot below, we can clearly see some outliers in the data, especially the two houses in the lower-right corner with living area greater than 4000 square feet and price lower than $200,000.

The age of the house also plays an important role in its price. Newer houses have higher average prices. There are several houses built before 1900 having a high price.



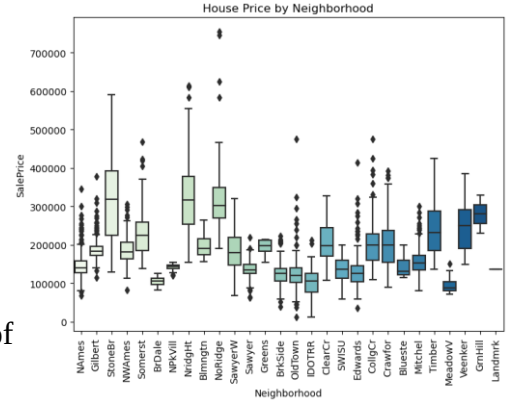
Surprisingly, houses with garage which can hold 4 cars are cheaper than houses with 3-car garage, but with the chart, we can see there are many outliners in house having 1 or 2 garages with the range of high fluctuation. For the number of bathrooms, it is clearly stated that house with 3 bathroom is also have the highest valuation in term of determining price of house

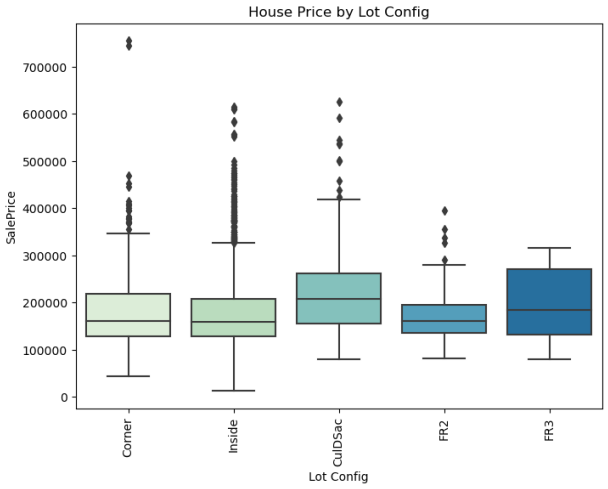




c. Categorical Variables Correlation with Sales Price

There is a big difference in house prices among neighborhood in Ames. The top 3 expensive neighborhoods are NridgHt, NoRidge and StoneBr with median house prices of approximately $300,000, three times as high as the median of the 3 cheapest neighborhoods, which are BrDale, DOTRR and MeadowV. Other seem no much different with multiple outliners, we can exclude this determinants out of factors in our models.





In term of the Lot Config, we can observe that there is no much different between different Log Config which is affected the houses price. So we shouldn't put the Lot Config into predicting the house prices.

3. MODEL ESTIMATION AND SELECTION

We select multiple linear regression model for prediction of house price as linear regression model is a fundamental tool that has various advantages over other regression models such as polynomial model and logistic model. Moreover, the model may be the most interpretable regression model available that has direct access to get the results

Initially, we will split the test and train datasets with a test size of 30% of total datasets. Then, since we would not like to predict based on a series of floated datasets thus, we want to fix it with a random state of 1. Then check the shape of train test variables to ensure shape of test and trained datasets are the same. When shape of train and test variables is fitted, rewrite the tests using trained model as parameters to fit linear regression model.

After data selection and fit in the linear regression, we reshape all independent factors then fit to linear regression again then print the coefficient of each individual variable. After candidate all the parameter, we will build the model based on external and internal factors of the data set. So we got 3 models can be written as:

Model 1: Internal Factors that affect the Log sales Price: Overall Quality, Full Bath, Garage Cars

$$LogSale\ Price = 10.56 + 0.17 * Overall\ Quality + 0.11 * Full\ Bath + 0.13 * Garage\ Car$$

$\beta_1 = 0.17$ means that an increase of 1% in Overall will result in 0.17% increase in house price when other predictors hold constant.

$\beta_2 = 0.11$ indicates that an increase of 1% in Full Bath will result in 0.11% increase in house price when other predictors hold constant.

$\beta_3 = 0.13$ indicates that an increase of 1% in Garage Car will result in 0.13% increase in house price when other predictors hold constant.

Model 2: External Factors that affect the Log sales Price: Year Build, Group Living Area

$$LogSale\ Price = -1.3 + 0.006 * \text{Year Build} + 0.0005 * \text{Group Living Area}$$

$\beta_1 = 0.006$ means that an increase of 1% in Year Build will result in 0.006% increase in house price when other predictors hold constant.

$\beta_2 = 0.0006$ indicates that an increase of 1% in Group Living Area will result in 0.0005% increase in house price when other predictors hold constant.

Model 3: Combine both external and internal factors to predict Log Sales Price

$$LogSale\ Price = 5.11 + 0.12 * Overall\ Quality - 0.03 * Full\ Bath + 0.08 * Garage\ Car + 0.003 * \text{Year Build} + 0.0003 * \text{Group Living Area}$$

$\beta_1 = 0.12$ means that an increase of 1% in Overall will result in 0.12% increase in house price when other predictors hold constant.

$\beta_2 = -0.03$ indicates that an increase of 1% in Full Bath will result in 0.03% decrease in house price when other predictors hold constant.

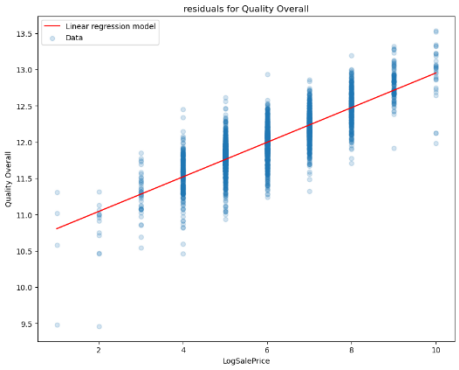
$\beta_3 = 0.08$ indicates that an increase of 1% in Garage Car will result in 0.08% increase in house price when other predictors hold constant.

$\beta_4 = 0.003$ means that an increase of 1% in Year Build will result in 0.003% increase in house price when other predictors hold constant.

$\beta_5 = 0.0003$ indicates that an increase of 1% in Group Living Area will result in 0.0003% increase in house price when other predictors hold constant.

<u>Check assumption of model</u>

We check linearity assumption of these two variables. We set maximum Overall Quality sold based on the house price as 100 and transparency of the data as 0.3. Then we get the residual plot of quantity.



The plot explicitly shows that there is a linear relationship between house price and Overall Quality.

4. MODEL EVALUATION

Now, we are going to evaluate the model. Using MSE (Mean Squared Error) is the arithmetic mean of sum of the squared predicted error. Predicted error is the difference between actual dependent variables and predicted dependent variables, which shows distance between the residual and linear regression line on the residual plot. MSE could be written as $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ where y_i is actual output and $\hat{y}_i$ is predicted output. In our case, the actual output is trained model, thus MSE can be written as:

$$MSE\ for\ Model\ 1 = \frac{1}{n}\sum_{i=1}^{n}(y_{train} - \hat{y}_i)^2 = 0.041122344509120674$$

$$MSE\ for\ Model\ 2 = \frac{1}{n}\sum_{i=1}^{n}(y_{train} - \hat{y}_i)^2 = 0.051853610062087635$$

$$MSE\ for\ Model\ 3 = \frac{1}{n}\sum_{i=1}^{n}(y_{train} - \hat{y}_i)^2 = 0.0322017971037722$$

For matching the unit of the output to the power of 1, we calculate root of MSE which is RMSE. RMSE brings the square of error back to the same level of prediction to ensure that the result has same unit of predicted output. However, we cannot interpret much more from a single number thus MSE (or RMSE) could only be compared the fitness with all 3 models.

Therefore, we use coefficient of determination (R^2) to assess the model. R^2 measures how much dependent variables can be explained by the independent variables. R^2 for our case is 0.75, 0.69 and 0.8 for each model respectively. which means that house price can be 75% explained by the regression model 1, 69% explained by the regression model 2 and 85% explained by the regression model 3. However, R^2 does not prevent overfitting of the data, thus use adjusted R^2 to take into consideration of overfitting. The result of adjusted R^2 is 0.75, 0.69 and 0,8, is much similar with the R^2 indicates that the model is pretty robust.

The shortcoming of the model is we do not test significance of each coefficient. We cannot ensure all coefficients are significantly differ to zero. Also, we do not do the normality test of the errors thus cannot ensure the model is unbiased. From the correlation test, it states that there is possible multicollinearity among predictors as the predictors' correlation are not zero means that the predictors in linear regression model correlated with each other. Existence of multicollinearity will make the linear regression problematic because we cannot hold other predictors constant when change one of predictors.

5. CONCLUSION

After conducted a detailed EDA and building 3 multiple linear regression models depending on both external and internal factors, we can conclude the combination of both internal and external factors when evaluating the price of the house. However, with the average coefficient of determination of 3 models, there should be more and more factors should be considered to make the comprehensive prediction of the house prices.