A statistical predictive analysis of the sale of the houses

Part 1: Business context and problem formulation

We are given 79 explanatory variables describing almost every aspect of residential homes in the US, gathered from 2006 to 2010.

The aim of this analysis is finding a model that can accurately predict the prices of the houses and defining the main factors that influence pricing. A developed casual and optimal model and defined significant factors will be the outcome of this report.

In this report, I am going to use a linear regression model, a kNN regression model, and a Random Forest Regressor model to predict the price to the houses. I will use mean absolute error to measure the accuracy of the models

.

Part 2: Data processing

i) Data cleaning

Before starting to analysis, it is necessary to have a outlook of the dataset. From a brief explore analysis we are clear that there are 1460 observations in the dataset with 36 independent numerical variables and 43 independent categorical variables.

To clean to data, the variables with more than 10 null entries are deleted and then the observations which still with null variable entries are deleted. This is because I want to keep as much original data as possible. And in this case fill the null entries with the mean value is not suitable because the big amount of missing data and in consideration of complicity and computation of time.
After the cleaning, there are 1451 observations in the dataset with 34 independent numerical variables and 29 independent categorical variables.

ii) Categorical variables processing

By observing, it is easy to see many remained categorical variables represent the rank of the house condition in different aspects. So, them can be convert to numerical variables to simplify. I imitate the ranking numerical variables that provided to do the convert, small numbers from 1 represent low quality big numbers below 10 represent high quality.

Here is the list of variables that are converted: ExterQual, ExterCond, HeatingQC, KitchenQual, LotShape, Utilities, LandSlope.

As for other categorical variables who represent the location, material, and some professional factors about the houses, in consideration of complicity and computation of time to analysis big number of categorical variables by plotting side by side boxplot one by one, I make a compromise to not use these variables.

iii) Splitting into train and test set

In the following models we will use the train set to train the models and the test set to test the ability of the prediction of the model.
But in the exploratory analysis for the target variable SalePrice part I choose to use the whole data set to have a whole view of the distribution of it.
we specify that the training set will contain 70% of the data. The random state parameter is 1. By using this random state, we will always get the same training and test set.

Part3: EDA

.

i) Univariate Exploratory Analysis

A. Response Variable: SalePrice

Exploratory analysis is conducted on the response variable MPG, recoded as variable SalePrice. The summary statistics of SalePrice is presented in the table below.

Table 3.1 Summary Statistics

	SalePrice
count	1451.0000
mean	180624.1020
std	79312.1283
min	34900.0000
25%	129900.0000
50%	162500.0000
75%	214000.0000
max	755000.0000

Table 3.2 Skewness and Kurtosis

	SalePrice
skewness	1.8811
kurtosis	6.5463

There are 1451 data points remained in total. The mean price of house is 180624.1020 units , standard deviation of the data is 79312.1283, with a range of 31. The median is 162500 units, lower than the mean, which suggests the data is right skewed. This is confirmed by Table 3.2, which shows that skewness has a value of 1.8811. The kurtosis value of the dataset from Table 3.2 is 6.5463.

A histogram and boxplot are drawn to further examine the distribution of SalePrice.

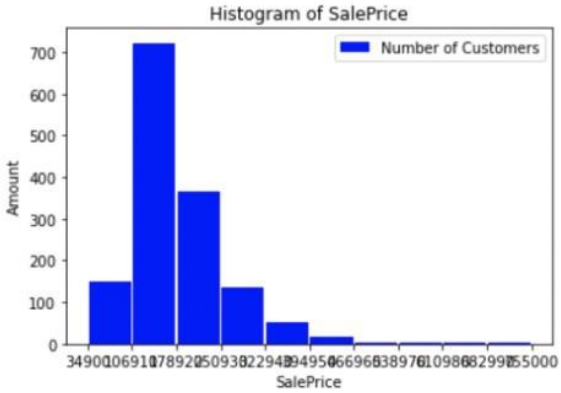


Figure 3.3 Histogram 'SalePrice'

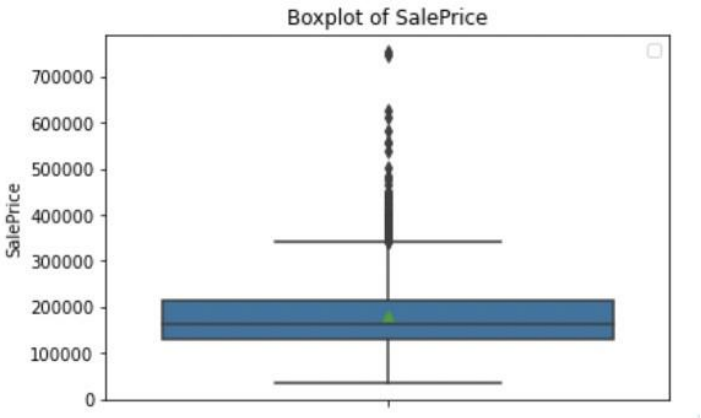


Figure 3.4 Boxplot 'SalePrice'

The two graphs confirm visually that the data is right skewed, and that there may be some outliers in the longer right tail of the distribution.

ii）Bivariate Exploratory Analysis

A. Primary Factor: OverallQual

By sorting out the correlation between independent variables and SalePrice, we found out OverallQual has the highest correlation with SalePrice. And the correlations of other highly related variables to SalePrice are shown in figure 3.5 below. The correlation coefficient between SalePrice and engine displacement has a value of 0.7963, suggesting a strong negative relationship between the two variables

OverallQual	0.796326
GrLivArea	0.747181
ExterQual	0.685097
KitchenQual	0.681633
TotalBsmtSF	0.657891
1stFlrSF	0.646456
GarageCars	0.639051
GarageArea	0.623788
TotRmsAbvGrd	0.553103
FullBath	0.550809

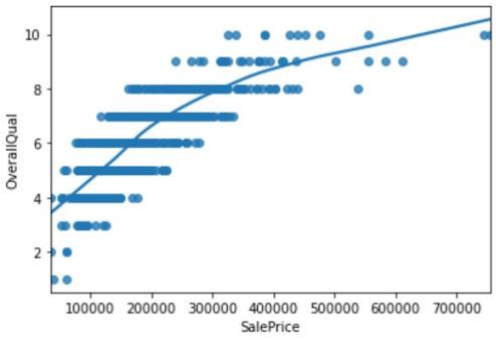


Figure 3.5 correlation with SalePrice of top ten variables Figure 3.6 OverallQual vs SalePrice scatter plot ”

To explore the relationship further, a scatter plot is drawn with a locally smooth regression line, shown in figure 3.6 above. The graph is consistent with the observation that a positive relationship exists between SalePrice and OverallQual. However, it also suggests that the negative relationship is not simple, as the data points resemble a non-linear pattern, and the locally smoothed regression line exhibits a similar curve.

B. Other Numerical Response Variables

To examine the relationship between predictors and SalePrice, the correlation heat map in figure 3.7 is examined.

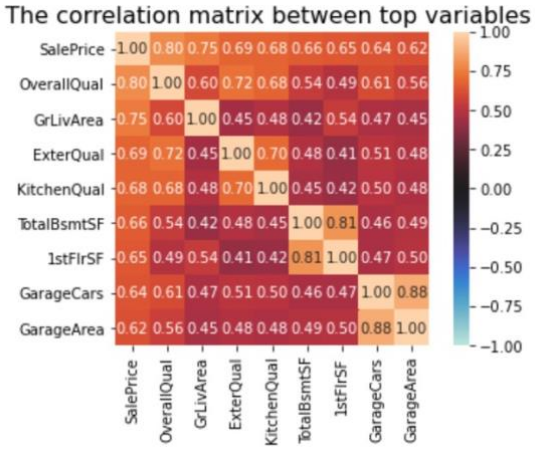


Figure 3.7 Variable Correlation Heat Map

Looking at the first row of the heat map, we can see the selected variables all has positive correlation with SalePrice. Table 3.7 also shows a positive correlations between predictors, but most of them are weak so we do not consider omitted variable bias.

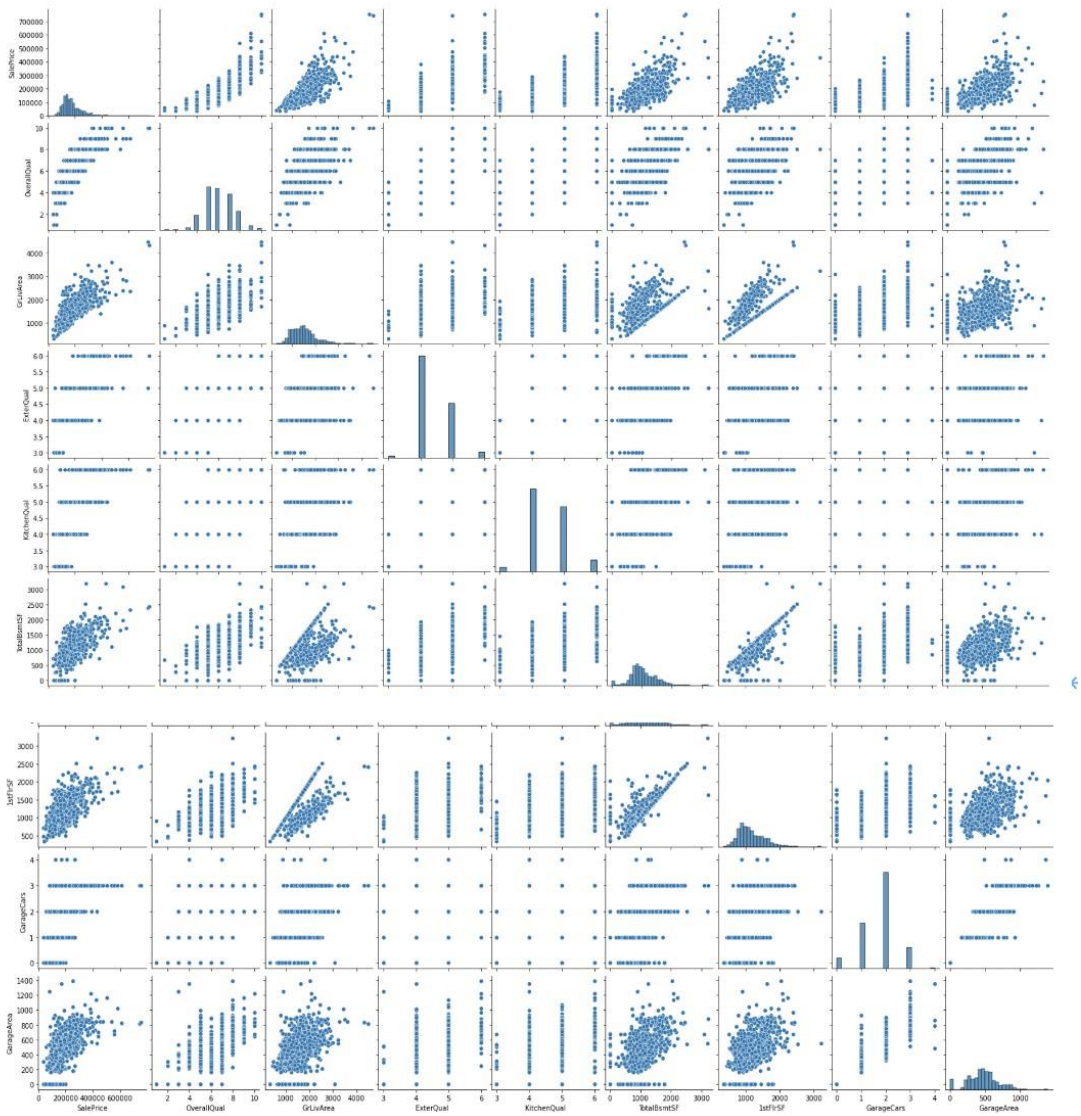


Figure 3.8 Scatter Plot Matrix

A scatter plot matrix is used to explore the linearity of predictor's relationship to MPG.
Most of the variables seem related to SalePrice in a positive relationship. And obviously there are many ranking variables which make the relationship with SalePrice is nonlinear.
It makes logical sense that when these ranks in quality are higher the house is in a better quality and higher price.
As expected from the correlation analysis, the predictors are also related to each other. A relatively strong positive linear relationship can be observed between GrLivArea and 1stFlrSF, GrLivArea and TotalBsmSf, as well as GarageCars and GarageArea. That makes logical sense because these pairs describe same traits of a house. This warns us to pay attention to multicollinearity later.

ii) List of potential variables

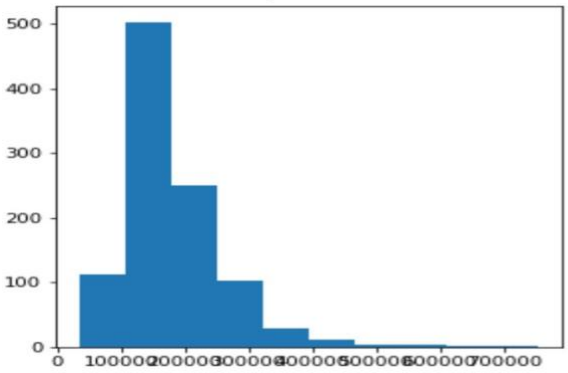
Therefore, the remaining variables are the ones that we believe to be the potential causes of SalePrice by correlation. The top 8 correlated variables are remained.
'SalePrice','OverallQual','GrLivArea','ExterQual','KitchenQual','TotalBsmtSF','1stFlrSF','GarageCars','Garage Area']]

Part4: Three models

4.1 Linear Regression

i) Normalize the target variable

From the EDA, we can see the distribution of the target variable SalePrice is obviously right skewed. Transform skewed numeric features using log(p+1) transformation makes the distribution of SalePrice more normal. In the linear regression, I will use the transformed SalePrice as the target variable.



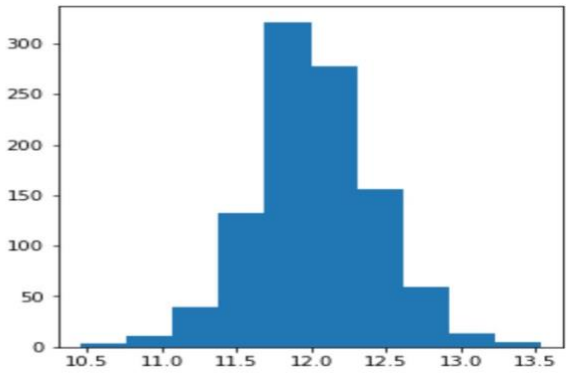


Figure 4.1 original distribution *4.2 log transformed distribution*

ii) Standardize the predictors

In our dataset, the predictors have quite different ranges. For GrLivArea, which describes above grade (ground) living area in square feet and has a mean of 1514.1743. But for other ranking variables the range is from 1 to 10. So we need to standardize the predictors in both train and test data. Means of the predictors in training data after standardization is more centered and have a zero mean and a 1 standard deviation. By doing this it is more numerically stable.

iii) Model with Variables from primary selection

The model formula : $\log(Y)=\beta 0 +\beta 1X1 +...+\beta 1Xp +\varepsilon$

Based on hypothesis testing, Adj. R-squared and AIC/BIC criteria, the variable selection is conducted.

With OLS algorithm to train our linear model, first, we construct a MLR with the primary selected variables and get :
Adj. R-squared: 0.8540, AIC: -904.0000, BIC: -859.7000.

The results including significant test in table below:

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.855
Model:	OLS	Adj. R-squared:	0.854
Method:	Least Squares	F-statistic:	741.0
Date:	Mon, 19 Sep 2022	Prob (F-statistic):	0.00
Time:	17:33:50	Log-Likelihood:	461.02
No. Observations:	1016	AIC:	-904.0
Df Residuals:	1007	BIC:	-859.7
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	288.8457	11.199	25.792	0.000	266.869	310.822
OverallQual	0.1812	0.012	15.082	0.000	0.158	0.205
GrLivArea	62.1769	3.396	18.310	0.000	55.513	68.841
ExterQual	0.0113	0.004	2.552	0.011	0.003	0.020
KitchenQual	0.0325	0.005	6.628	0.000	0.023	0.042
TotalBsmtSF	33.7496	3.808	8.862	0.000	26.277	41.223
1stFlrSF	-0.8731	3.436	-0.254	0.799	-7.616	5.870
GarageCars	0.0356	0.008	4.374	0.000	0.020	0.052
GarageArea	5.0218	2.268	2.214	0.027	0.571	9.473

Table 4.3 Results Summary: MLR with primary selected variables

The table suggests that EterQual,1stFlrSF, and GarageArea are not significant for the P-value is 0.0110,0.7990,and 0.270 respectively, which are greater than our manually chosen alpha of 0.01. So, we decided not to include them in the MLR.

iv) Model after reducing variables

The new MLR after reducing variables and get:
Adj. R-squared:0.8520, AIC: -898.3000, BIC: -868.8000
The minor decrease in Adj. R-squared is tolerable. The minor increase in AIC is tolerable.
And we get BIC smaller which is good for the model.
The results of the MLR are shown in the table below after reducing the insignificant variables.

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.853
Model:	OLS	Adj. R-squared:	0.852
Method:	Least Squares	F-statistic:	1173.
Date:	Mon, 19 Sep 2022	Prob (F-statistic):	0.00
Time:	17:34:21	Log-Likelihood:	455.17
No. Observations:	1016	AIC:	-898.3
Df Residuals:	1010	BIC:	-868.8
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	281.7804	10.692	26.353	0.000	260.799	302.762
OverallQual	0.1919	0.011	17.231	0.000	0.170	0.214
GrLivArea	61.8368	3.156	19.592	0.000	55.643	68.030
KitchenQual	0.0378	0.005	8.335	0.000	0.029	0.047
TotalBsmtSF	34.6015	2.524	13.710	0.000	29.649	39.554
GarageCars	0.0511	0.005	10.608	0.000	0.042	0.061

Table 4.4 Results Summary: MLR after reducing

After deleting, the MLR suggests these factors are all positively related to MPG, holding the other variable constant. And we also notice that the coefficient of the intercept is large and the coefficient for the variables is small, which suggests in general the SalePrice does not change much among different conditions.
The effect of these variables is summarized as:

Keeping all other variables constant, a 1 unit increase in OverallQual is associated with an approximate 0.0019% typical change in SalePrice.
Keeping all other variables constant, a 1 unit increase in GrLivArea is associated with an approximate 0.6184% typical change in SalePrice.
Keeping all other variables constant, a 1 unit increase in KitchenQual is associated with an approximate 0.0004% typical change in SalePrice.
Keeping all other variables constant, a 1 unit increase in TotalBsmSF is associated with an approximate 0.3460% typical change in SalePrice.
Keeping all other variables constant, a 1 unit increase in GarageCars is associated with an approximate 0.0005% typical change in SalePrice.

The MAE of linear regression model is 182310.2758.

4.2 kNN model

In the EDA we learned that the distribution of SalePrice is right skewed. And we learned that OverallQual has the highest correlation with SalePrice. So we choose OverallQual to predict in kNN model.
In kNN model, first we manually choose k=10.

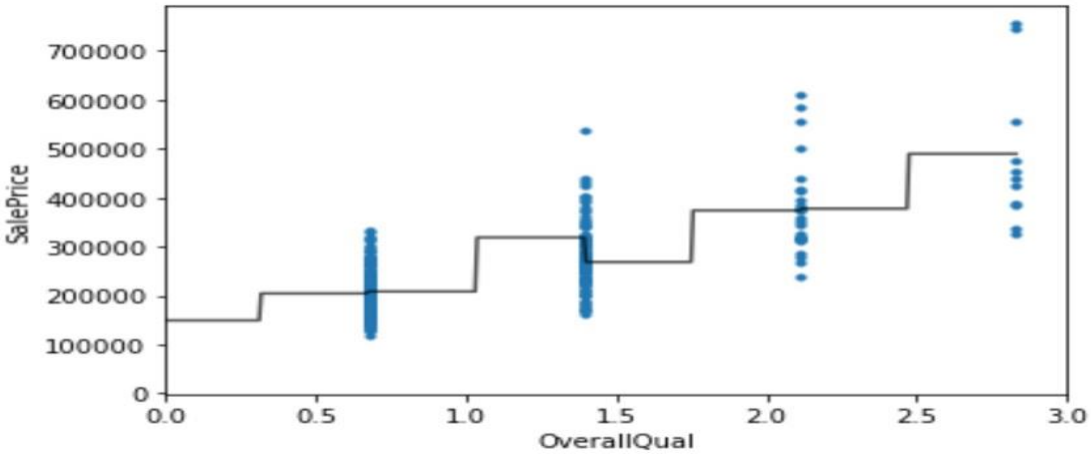


Table 4.5 kNN with k=10

From the plot we can see the prediction line has low ability to align with the real value.

i) Variable selection

To fix it, first we add other predictors, by the order of the highest correlation with SalePrice.
As the experiments result suggest we add GrLivArea and ExterQual.
Because when we add more predictors the MAE would not keep decreasing.
the MAE results are as follows:
when predictors is OverallQual, the MAE is 187.3190
when predictors are OverallQual and GrLivArea, the MAE is 165.9357
when predictors are OverallQual, GrLivArea and ExterQual, the MAE is 160.39786 and since then adding predictor will not make the MAE decrease.

ii) Value of k

Second we change the value of k to find the suitable value of k to make the model has the lowest MAE.

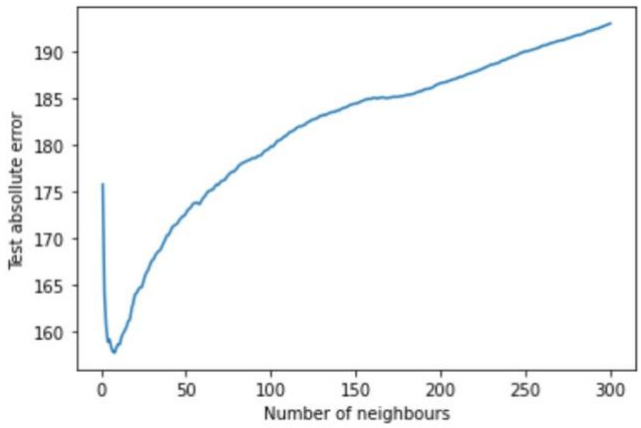
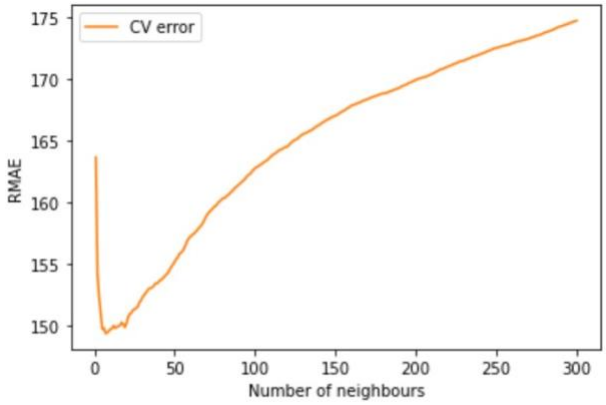


Table 4.6 how k changes the MAE by simply changing k

As table 4.6 suggests, we know when k=8 the MAE is the lowest.
Also we can use cross-validation to find a best value of k.




```
Lowest CV error: K = 7
```

Table 4.6 how k changes the MAE by simply changing k cross-validation

As the cross-validation method suggests k=7.
Then we try to set k=7 get MAE= 157.8158,
set k=8 get MAE= 157.6861
the MAE is lower when k=8
So we can form a kNN model with predictors of OverallQual, GrLivArea and ExterQual and k=8.

4.3 Random Forest model

Code reference : https://github.com/data-doctors/kaggle-house-prices-advancedregression-techniques/blob/master/04-modelling/04-modelling.ipynb

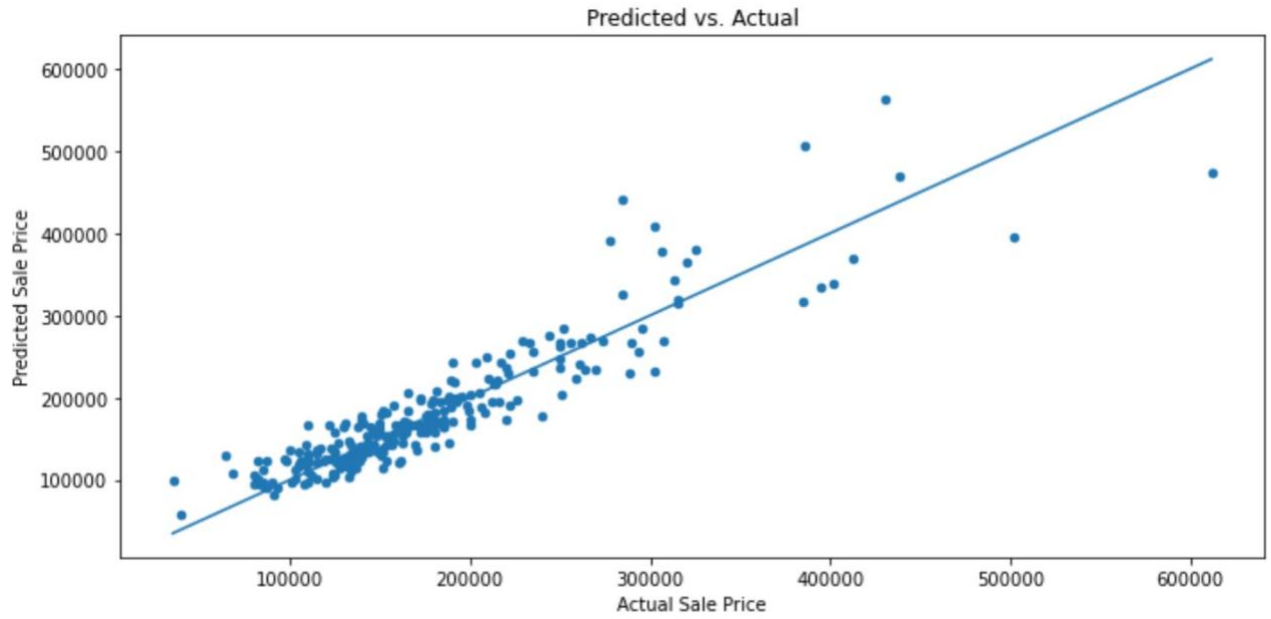


Table 4.7 Random forest model prediction vs actual value

The MAE of Random Forest model is 20041.2190.

Part 5: Conclusion

5.1 final model

The MAE of linear regression model is 182310.2758.
The MAE of kNN model is 157.6861
The MAE of RandomForest model is 20041.2190.
So we choose kNN as the optimal model.

5.2 significant factors

By linear model suggests, OverallQual, GrLivArea, KitchenQual, TotalBsmSF and GarageCars are the significant factors.
By kNN model suggests, OverallQual, GrLivArea and ExterQual are the significant factors.

5.3 Overall conclusion

The overall goal of this study is to identify the most effective factors for developing a causal model.
To summarise the analysis process, we first investigated all factors in the given data and used the most relevant ones to build various models.Then I compare models by comparing the MAE of the models.
We are aiming to offer model with lowest MAE.

With the significant factors mainly describing the quality rank and the area of the house, we can say the main factors that influence the house price is about the area and the quality. But we also have some limitation of this analysis. This report did not use the data related to the location and constructive material of the house, which is generally believed every import to the houses price. In order to aim a simple and effective model we compromise in not using these complex data but we should be aware that these factors can have significant influence of the house price. **1. single predictor model**

kNN is the best predictive model that uses a single predictor as the previous result shows:
when predictors is OverallQual, the MAE is 187.3190

when predictors are OverallQual and GrLivArea, the MAE is 165.9357, which means the predictive ability of single predictor model in kNN is not far away lower than the multiple predictor models.

2. *median* error

Yes I may change my method. The target variable SalePrice is distributed right skewed.

The outliers has more influence to the mean than the median. In this report, I have use transformed target variable as the respond. But if we use median error as the standard I may not use transform to simplify the linear model.

3.Predict sale prices for three houses for the year 2022.

With the predictor of OverallQual =5, in kNN model the price is $489548 in 2006-2010.
We choose 1.75%(https://zh.tradingeconomics.com/united-states/interest-rate)as the interest rate. We assume the data is from 2008 as the mean of 2006 and 2010.
OverallQual=5,P1=489548*(1+0.0175)^14=$624132.979
OverallQual=2,P2= =373861*(1+0.0175)^14=$ 476641.677
OverallQual=1,P2= = 208285*(1+0.0175)^14=$ 265546.05