Predicting the Sale Price of the House using Multiple Linear Regression

Yash Bhargava
Dept. of Computing
National College of Ireland
Dublin, Ireland
x22220861@student.ncirl.ie

I. Introduction

This task involves the analysis of the housing dataset by identifying the factors that are influencing the sale price of the house and checking the relationship between the independent features and the dependent variable, i.e., the sale price of the house, using Multiple linear regression. **Multiple Linear Regression:** It is a regression-based statistical model that is used in the prediction of a continuous variable in which there is a dependent variable, which is denoted as the y variable, and two or more independent variables that are independent of each other, denoted as the x variable, and there is a linear relationship between the y and x variables. The equation of the linear regression is shown as:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\epsilon}_i \quad (1)$$

It consists of various types of categorical and continuous variables, namely Lot Frontage, Lot Area, Building Type, Building Style, Overall Condition, Year of Built External Condition, Total Basement Area, First Floor Area, Second Floor Area, Full Bathrooms, Half Bathrooms, Bedroom Above Ground, Kitchen Above Ground, Fireplaces, Longitude, Latitude, and Sale Price of House

II. Dataset Description

This data set includes 2413 observations and 18 columns, in which there are 17 independent variables and 1 independent variable, i.e., sale price (y), for which the description of all the variables is given below:

- **Lot_Frontage:** The area of the street in front of the house
- **Lot_Area:** The area of the plot is in square feet.
- **Bldg_Type:** The type of house
- **House_Style:** The style of the house
- **Overall_Cond:** Overall condition of the house.
- **Year_Built:** Year of construction.
- **Exter_Cond:** Condition of the house from the exterior.
- **Total_Bsmt_SF:** Basement area in square feet
- **First_Flr_SF:** Area of the ground floor in square feet
- **Second_Flr_SF:** Area of the first floor in square feet
- **Full_Bath:** Count of full bathrooms.
- **Half_Bath:** Count of half bathrooms.
- **Bedroom_AbvGr:** Count of bedrooms on or above the ground floor.
- **Kitchen_AbvGr:** Count of kitchens on or above the ground floor.
- **Fireplaces:** Count of fireplaces.
- **Longitude:** Longitude of the house.
- **Latitude:** Latitude of the house.
- **Sale_Price:** Selling price of the house (dependent variable y).

III. Exploratory Data Analysis

In this part of the analysis, I have segregated the data based on the variables, i.e., qualitative and quantitative, and then further divided the qualitative data into nominal and ordinal variables and the quantitative data into discrete and continuous variables to analyze the data easily and efficiently and apply the exploratory data analysis steps to each type of data.

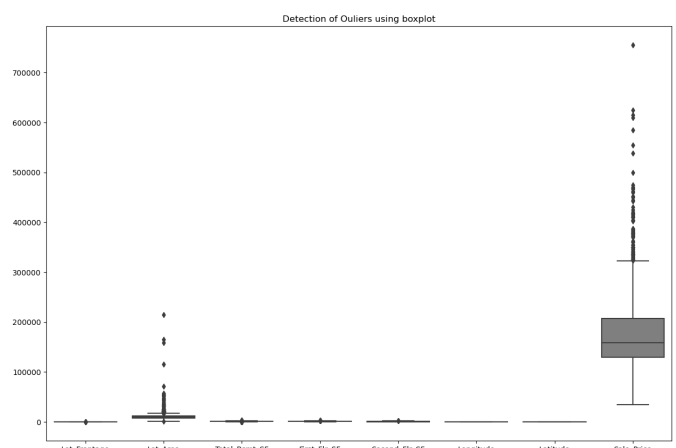


Fig. 1. Detection of Outliers

Secondly, in this part, I detected the missing values using the.info() function, and I found no missing values in this dataset and created a boxplot to detect the outliers. This author has found that there are outliers in various variables, as shown in Fig. 1. In qualitative data, I have described the data using the describe() function, and this author has a summary for the categorical variables for which I have the count, unique, top, and frequency. I have determined that the author has four categorical variables, of which there are two nominal variables and two ordinal variables.

IV. Data Preparation

In this part, I have started with the treatment of outliers, which I have done using the clip method, replacing the outliers with the 1subst percentile and 99th percentile and again checking whether the outliers are treated and plotted on the boxplot as shown in Fig. 2

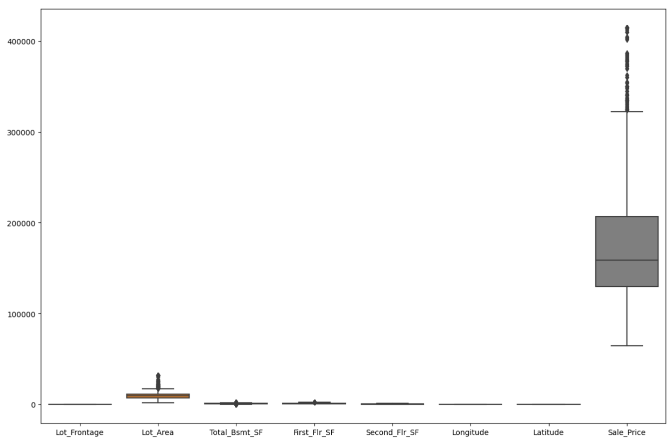
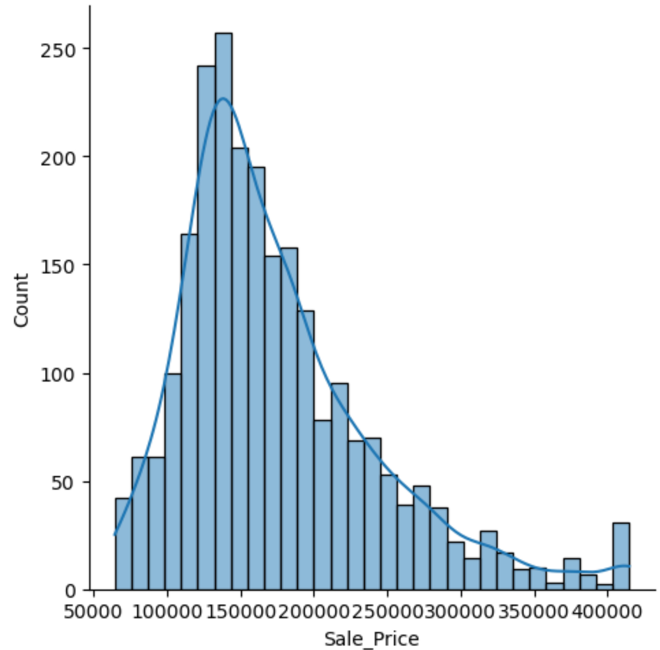


Fig. 2. After Treatment of Outliers



centring

Fig. 3. Distribution of Sale Price

Then this author checked those assumptions if this author was working on regression problems, which are the following:

1) All the dependent variables (y) and independent variables (x) must follow the normal distribution. If not all, at least the y variable must follow the normal distribution.
2) Y variable should be linearly related to the x variables; otherwise, we will not get the best-fit line.
3) The number of observations should be greater than the number of variables.
4) X should be independent variables.

Now we proceed to check the distribution of the Sale Price variable and we conclude that the skewness of this variable is 1.22 as well as that it is positively skewed, as shown in Fig. 3, and for the variable to be normally distributed, its skewness should be 0.

So, I have transformed the Sale Price (y) variable, taking the log n values for that variable and then considering them as the actual values of our dependent variable y. After transforming the y variable, I again checked the distribution, and then we got a normally distributed curve as shown in Fig. 4 and a skewness of 0.16, which is not exactly 0, but yes, we can say it is near 0.

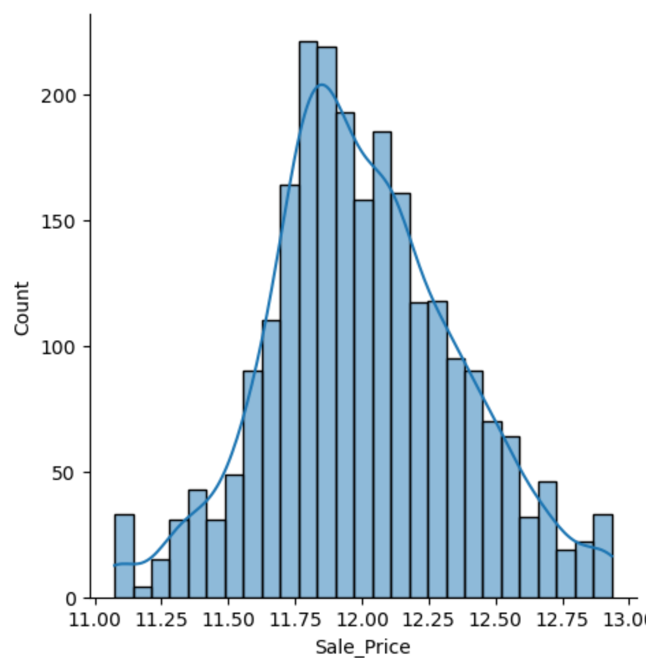


Fig. 4. Distribution of Transformed Sale Price

Now we have done our data preparation part on our continuous columns and further proceeded to handle our categorical variables, and I have created dummy variables so that we can build a good model on this dataset. But before building it, we combined our qualitative and quantitative data using the pandas.concat() function into a new data frame, and on this combined data frame, we will perform further steps. So now I have created a new transformed independent variable in our new data frame, which is housing_new, and stored the log values in this variable that would act as actual values of our dependent variable (y).

Then we checked the co-relation of this dataset and found that we have a positive intermediate co-relation with "First_Flr_SF" and "Total_Bsmt_SF" variables, i.e., 0.63 and 0.64, respectively. I have shown that with the following heatmap in Fig. 5, as shown:

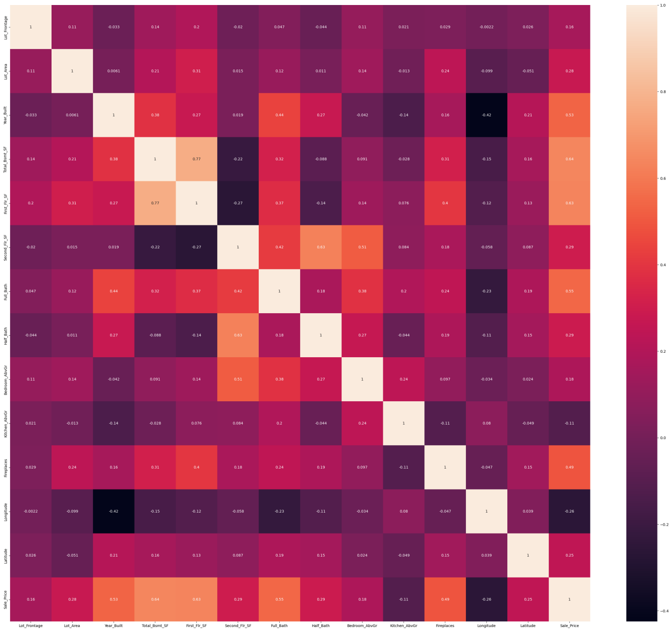


Fig. 5. Correlation between Sale Price and other variables

V. Modelling

So, here in this phase of modelling, I have followed all the data modelling steps, including the following steps:

Splitting the Data into Training and Test Set

I have divided the data into the training set and test set with the ratio of 80:20, i.e., 80%

Model Building

Model 0: Considering All Features: In this model 0, I have considered all the variables and initiated with the following:

Definition of the Model: In this, I have defined the model by applying the linear regression formula and considering all the variables. The model equation is shown in Fig. 6

```
'ln_Sale_Price ~ Bedroom_AbvGr+Bldg_Type_OneFam+Bldg_Type_Twnhs+Bldg_Type_TwnhsE+Bldg_Type_TwoFmCon+Exter_Cond_Fair+Exter_Cond_
Good+Exter_Cond_Poor+Exter_Cond_Typical+Fireplaces+First_Flr_SF+Full_Bath+Half_Bath+House_Style_One_and_Half_Fin+House_Style_On
e_and_Half_Unf+House_Style_SFoyer+House_Style_SLvl+House_Style_Two_Story+House_Style_Two_and_Half_Fin+House_Style_Two_and_Half_
Unf+Kitchen_AbvGr+Latitude+Longitude+Lot_Area+Lot_Frontage+Overall_Cond_Average+Overall_Cond_Below_Average+Overall_Cond_Excelle
nt+Overall_Cond_Fair+Overall_Cond_Good+Overall_Cond_Poor+Overall_Cond_Very_Good+Overall_Cond_Very_Poor+Second_Flr_SF+Total_Bsmt
_SF+Year_Built'
```

Fig. 6. Model 0 Summary

Fitting the Model: After the definition of the model, we fit the model using the fit() function.

Summary of the Model: I got the summary of the model using the summary method and found out that we got multiple R2 values of 0.885 and an adjusted R2 squared value of 0.883, as shown in the model summary in Fig. 7, and some insignificant features are overfitting our model, affecting its accuracy.

Variable Reduction

This step of modelling is done by doing the following two steps:

```
==============================================================================
Dep. Variable:            ln_Sale_Price   R-squared:                       0.885
Model:                              OLS   Adj. R-squared:                  0.883
Method:                   Least Squares   F-statistic:                     404.6
Date:                  Thu, 23 Nov 2023   Prob (F-statistic):               0.00
Time:                          11:03:17   Log-Likelihood:                 1317.6
No. Observations:                  1930   AIC:                            -2561.
Df Residuals:                      1893   BIC:                            -2355.
Df Model:                            36
Covariance Type:              nonrobust
===============================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
Intercept                     -50.9132     14.503     -3.511      0.000     -79.356     -22.471
Bedroom_AbvGr                  -0.0461      0.005     -9.535      0.000      -0.056      -0.037
Bldg_Type_OneFam                0.0898      0.028      3.239      0.001       0.035       0.144
Bldg_Type_Twnhs                -0.0060      0.032     -0.186      0.852      -0.069       0.057
Bldg_Type_TwnhsE                0.0945      0.030      3.114      0.002       0.035       0.154
Bldg_Type_TwoFmCon              0.0653      0.028      2.318      0.021       0.010       0.121
Exter_Cond_Fair                -0.0460      0.052     -0.879      0.379      -0.149       0.057
Exter_Cond_Good                -0.0137      0.048     -0.288      0.774      -0.107       0.080
Exter_Cond_Poor                 0.0830      0.149      0.556      0.578      -0.210       0.375
Exter_Cond_Typical             -0.0099      0.048     -0.206      0.837      -0.104       0.084
Fireplaces                      0.0544      0.005     10.271      0.000       0.044       0.065
First_Flr_SF                    0.0004   1.68e-05     23.645      0.000       0.000       0.000
Full_Bath                       0.0276      0.008      3.314      0.001       0.011       0.044
Half_Bath                       0.0039      0.009      0.440      0.660      -0.013       0.021
House_Style_One_and_Half_Fin    0.0153      0.015      1.043      0.297      -0.013       0.044
House_Style_One_and_Half_Unf    0.0104      0.032      0.326      0.745      -0.052       0.073
House_Style_SFoyer              0.0616      0.019      3.225      0.001       0.024       0.099
House_Style_SLvl                0.0338      0.015      2.284      0.022       0.005       0.063
House_Style_Two_Story           0.0257      0.018      1.439      0.150      -0.009       0.061
House_Style_Two_and_Half_Fin    0.1328      0.055      2.403      0.016       0.024       0.241
House_Style_Two_and_Half_Unf    0.0957      0.036      2.650      0.008       0.025       0.166
Kitchen_AbvGr                  -0.0889      0.026     -3.406      0.001      -0.140      -0.038
Latitude                        0.5029      0.171      2.949      0.003       0.168       0.837
Longitude                      -0.3418      0.125     -2.737      0.006      -0.587      -0.097
Lot_Area                     7.207e-06   8.41e-07      8.574      0.000    5.56e-06    8.86e-06
Lot_Frontage                    0.0002   9.18e-05      2.247      0.025    2.63e-05       0.000
Overall_Cond_Average           -0.0198      0.008     -2.339      0.019      -0.036      -0.003
Overall_Cond_Below_Average     -0.1351      0.017     -7.977      0.000      -0.168      -0.102
Overall_Cond_Excellent          0.2458      0.026      9.542      0.000       0.195       0.296
Overall_Cond_Fair              -0.2875      0.026    -11.123      0.000      -0.338      -0.237
Overall_Cond_Good               0.0905      0.010      9.058      0.000       0.071       0.110
Overall_Cond_Poor              -0.3797      0.063     -6.047      0.000      -0.503      -0.257
Overall_Cond_Very_Good          0.1370      0.014      9.857      0.000       0.110       0.164
Overall_Cond_Very_Poor         -0.3679      0.090     -4.069      0.000      -0.545      -0.191
Second_Flr_SF                   0.0004   2.03e-05     18.517      0.000       0.000       0.000
Total_Bsmt_SF                   0.0002   1.24e-05     18.105      0.000       0.000       0.000
Year_Built                      0.0045      0.000     25.651      0.000       0.004       0.005
==============================================================================
Omnibus:                       50.903   Durbin-Watson:                   1.999
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              119.154
Skew:                           0.024   Prob(JB):                     1.34e-26
Kurtosis:                       4.216   Cond. No.                     5.66e+07
```

Fig. 7. Model 0 Summary

Feature Selection: In this step of feature selection, I have done a bi-variate analysis, i.e. firstly I calculated the f-score and p-value of all the features using the f_regression method from sklearn.feature_selection and then based on that I dropped those variables which have p-values greater than 0.05 because considering those features which have higher p-value means that they did not affect the dependent variable i.e., Sale Price too much, hence they are considered as insignificant variables.

Multicollinearity: Now after removing those variables with higher p-values greater than 0.05, we have done a multi-collinearity check i.e., all the predictor variables are not linearly related to each other, and we checked this by calculating the variance inflation factor (VIF) and we do this using variance_inflation_factor method and then remove those features which have VIF value ¿ 5, but we don't do this in one go and do it one by one, e.g., we have feature "Bldg_Type_TwnhsE" which we create as dummy variable have largest VIF among all the other features which is equal to 8.878847e+00 ¿¿ 5, so I have dropped this variable, hence we remove the variables with high VIF value one by one as it affects the VIF values of other features.
After removing all the features that have a VIF value greater than 5 we are left with those features which we are going to use to build our next model.

Model 1: After Variable Reduction

I performed all the model-building steps again after doing variable reduction, and after fitting the model, we got the model summary of Model 1, and we found that the R2 squared value and adjusted R2 score value are 0.853 and 0.851, as shown in Fig. 8 respectively. Now we can see that there are some features for which their p-values are greater than 0.05. So, we will remove those features as well to improve the accuracy measures, as they are not that relevant for the prediction of the dependent variable, i.e., sale price. Hence, we will make another model, which is Model 2, without considering those features.

OLS Regression Results

Dep. Variable:	ln_Sale_Price	R-squared:	0.853
Model:	OLS	Adj. R-squared:	0.851
Method:	Least Squares	F-statistic:	380.7
Date:	Thu, 23 Nov 2023	Prob (F-statistic):	0.00
Time:	11:03:20	Log-Likelihood:	1082.0
No. Observations:	1930	AIC:	-2104.
Df Residuals:	1900	BIC:	-1937.
Df Model:	29		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-51.4548	16.251	-3.166	0.002	-83.326	-19.583
Bedroom_AbvGr	-0.0228	0.005	-4.461	0.000	-0.033	-0.013
Bldg_Type_OneFam	0.0255	0.013	1.954	0.051	-9.71e-05	0.051
Bldg_Type_Twnhs	-0.1023	0.020	-5.206	0.000	-0.141	-0.064
Bldg_Type_TwoFmCon	0.0417	0.026	1.608	0.108	-0.009	0.092
Exter_Cond_Good	0.0076	0.011	0.680	0.497	-0.014	0.030
Fireplaces	0.0752	0.006	12.828	0.000	0.064	0.087
First_Flr_SF	0.0003	1.87e-05	18.456	0.000	0.000	0.000
Full_Bath	0.0932	0.009	10.662	0.000	0.076	0.110
Half_Bath	0.0629	0.009	6.751	0.000	0.045	0.081
House_Style_One_and_Half_Fin	0.1442	0.013	11.496	0.000	0.120	0.169
House_Style_One_and_Half_Unf	0.0104	0.036	0.291	0.771	-0.060	0.080
House_Style_SFoyer	0.0788	0.021	3.800	0.000	0.038	0.119
House_Style_Two_Story	0.2292	0.013	18.264	0.000	0.205	0.254
House_Style_Two_and_Half_Fin	0.3616	0.059	6.086	0.000	0.245	0.478
Kitchen_AbvGr	-0.1418	0.020	-6.934	0.000	-0.182	-0.102
Latitude	0.5165	0.191	2.702	0.007	0.142	0.891
Longitude	-0.3532	0.140	-2.517	0.012	-0.628	-0.078
Lot_Area	7.017e-06	9.32e-07	7.529	0.000	5.19e-06	8.85e-06
Lot_Frontage	0.0003	0.000	2.621	0.009	6.8e-05	0.000
Overall_Cond_Average	-0.0071	0.009	-0.745	0.456	-0.026	0.012
Overall_Cond_Below_Average	-0.1438	0.019	-7.594	0.000	-0.181	-0.107
Overall_Cond_Excellent	0.2649	0.027	9.836	0.000	0.212	0.318
Overall_Cond_Fair	-0.3220	0.029	-11.286	0.000	-0.378	-0.266
Overall_Cond_Good	0.0986	0.011	8.768	0.000	0.077	0.121
Overall_Cond_Poor	-0.3479	0.063	-5.496	0.000	-0.472	-0.224
Overall_Cond_Very_Good	0.1299	0.016	8.309	0.000	0.099	0.161
Overall_Cond_Very_Poor	-0.2403	0.100	-2.401	0.016	-0.436	-0.044
Total_Bsmt_SF	0.0002	1.38e-05	15.767	0.000	0.000	0.000
Year_Built	0.0040	0.000	20.862	0.000	0.004	0.004

Omnibus:	61.282	Durbin-Watson:	2.005
Prob(Omnibus):	0.000	Jarque-Bera (JB):	110.312
Skew:	0.245	Prob(JB):	1.11e-24
Kurtosis:	4.064	Cond. No.	5.62e+07

Fig. 8. Summary of Model 1

Model 2: After removing a few more features based on p-value

So, after removing the features i.e., “House_Style_One_and_Half_Unf”, “Exter_Cond_Good”, “Overall_Cond_Average”, and “Bldg_Type_TwoFmCon” and “Bldg_Type_TwoFmCon”, I have followed all the steps to build the model 2 with all significant features that are left. But in the summary of this model, we can see in Fig. 8 that there is no difference in the R2 Score and Adjusted R2 Score, but there is one variable, i.e., “Bldg_Type_OneFam” which has a p-value of 0.05 in model 1 summary, and now in the summary of this model, the p-value of this variable was increased, hence we need to remove this variable as its p-value had increased to 0.11 as shown in Fig. 9, then the declared significance level, which is 0.05, and build another model.

OLS Regression Results

Dep. Variable:	ln_Sale_Price	R-squared:	0.853
Model:	OLS	Adj. R-squared:	0.851
Method:	Least Squares	F-statistic:	441.5
Date:	Thu, 23 Nov 2023	Prob (F-statistic):	0.00
Time:	11:03:21	Log-Likelihood:	1080.0
No. Observations:	1930	AIC:	-2108.
Df Residuals:	1904	BIC:	-1963.
Df Model:	25		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-51.6361	16.151	-3.197	0.001	-83.311	-19.961
Bedroom_AbvGr	-0.0220	0.005	-4.341	0.000	-0.032	-0.012
Bldg_Type_OneFam	0.0191	0.012	1.574	0.116	-0.005	0.043
Bldg_Type_Twnhs	-0.1057	0.019	-5.429	0.000	-0.144	-0.068
Fireplaces	0.0756	0.006	12.913	0.000	0.064	0.087
First_Flr_SF	0.0003	1.86e-05	18.381	0.000	0.000	0.000
Full_Bath	0.0927	0.009	10.646	0.000	0.076	0.110
Half_Bath	0.0626	0.009	6.730	0.000	0.044	0.081
House_Style_One_and_Half_Fin	0.1442	0.012	11.565	0.000	0.120	0.169
House_Style_SFoyer	0.0793	0.021	3.834	0.000	0.039	0.120
House_Style_Two_Story	0.2283	0.012	18.311	0.000	0.204	0.253
House_Style_Two_and_Half_Fin	0.3548	0.059	5.985	0.000	0.239	0.471
Kitchen_AbvGr	-0.1393	0.020	-6.838	0.000	-0.179	-0.099
Latitude	0.5100	0.191	2.671	0.008	0.136	0.884
Longitude	-0.3598	0.139	-2.582	0.010	-0.633	-0.086
Lot_Area	7.269e-06	9.17e-07	7.928	0.000	5.47e-06	9.07e-06
Lot_Frontage	0.0003	0.000	2.728	0.006	7.88e-05	0.000
Overall_Cond_Below_Average	-0.1393	0.018	-7.665	0.000	-0.175	-0.104
Overall_Cond_Excellent	0.2722	0.026	10.544	0.000	0.222	0.323
Overall_Cond_Fair	-0.3184	0.028	-11.345	0.000	-0.373	-0.263
Overall_Cond_Good	0.1035	0.010	10.420	0.000	0.084	0.123
Overall_Cond_Poor	-0.3489	0.063	-5.539	0.000	-0.472	-0.225
Overall_Cond_Very_Good	0.1360	0.014	9.427	0.000	0.108	0.164
Overall_Cond_Very_Poor	-0.2430	0.100	-2.431	0.015	-0.439	-0.047
Total_Bsmt_SF	0.0002	1.37e-05	15.900	0.000	0.000	0.000
Year_Built	0.0039	0.000	21.260	0.000	0.004	0.004

Omnibus:	58.347	Durbin-Watson:	2.004
Prob(Omnibus):	0.000	Jarque-Bera (JB):	105.129
Skew:	0.231	Prob(JB):	1.48e-23
Kurtosis:	4.046	Cond. No.	5.58e+07

Fig. 9. Summary of Model 2

Model 3: After the removal of all the insignificant variables

So, this is our final model after the reduction of the “Bldg_Type_OneFam” variable. we have built this model as our final model with all the significant features that are important for the prediction of the dependent variable y, i.e., Sale Price and also its R2 score and adjusted R2, which are the same as the previous model i.e., 0.853 and 0.851 respectively, as we can see in Fig. 11.

VI. Interpretation

Model 3, which is the final model considering all the significant features, has a p-value ¡=0.05. There is no multicollinearity between these features as their VIF ¡ 5, and we have a much better model than the previous ones. The formula for the Linear Regression model is shown in Fig. 11

```
'ln_Sale_Price ~ Bedroom_AbvGr+Bldg_Type_Twnhs+Fireplaces+First_Flr_SF+Full_Bath+Half_Bath+House_Style_One_and_Half_Fin+House_S
tyle_SFoyer+House_Style_Two_Story+House_Style_Two_and_Half_Fin+Kitchen_AbvGr+Latitude+Longitude+Lot_Area+Lot_Frontage+Overall_C
ond_Below_Average+Overall_Cond_Excellent+Overall_Cond_Fair+Overall_Cond_Good+Overall_Cond_Poor+Overall_Cond_Very_Good+Overall_C
ond Very Poor+Total Bsmt SF+Year Built'
```

Fig. 10.

The summary of the final model is represented in Fig. 11.

The coefficients of all the variables in this model indicate that if the value of the coefficient is positive, that means that with an increase of the value of that independent variable by one unit, the value of the dependent variables increases with the coefficient holding the other variables constant. If the coefficient of an independent variable is negative, it means that by keeping other variables constant, the predicted variable decreases by the coefficient of that variable. So, here in the final model, we can say that if an independent variable “fireplaces” has a coefficient value of 0.0755, which is a

```
                             OLS Regression Results
==============================================================================
Dep. Variable:           ln_Sale_Price   R-squared:                       0.853
Model:                             OLS   Adj. R-squared:                  0.851
Method:                  Least Squares   F-statistic:                     459.5
Date:                 Thu, 23 Nov 2023   Prob (F-statistic):               0.00
Time:                         11:03:21   Log-Likelihood:                 1078.8
No. Observations:                 1930   AIC:                            -2108.
Df Residuals:                     1905   BIC:                            -1968.
Df Model:                           24
Covariance Type:             nonrobust
===============================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
Intercept                     -52.8765     16.138     -3.277      0.001     -84.526     -21.227
Bedroom_AbvGr                  -0.0201      0.005     -4.082      0.000      -0.030      -0.010
Bldg_Type_Twnhs                -0.1176      0.018     -6.542      0.000      -0.153      -0.082
Fireplaces                      0.0755      0.006     12.897      0.000       0.064       0.087
First_Flr_SF                    0.0003   1.86e-05     18.310      0.000       0.000       0.000
Full_Bath                       0.0921      0.009     10.580      0.000       0.075       0.109
Half_Bath                       0.0629      0.009      6.760      0.000       0.045       0.081
House_Style_One_and_Half_Fin    0.1436      0.012     11.516      0.000       0.119       0.168
House_Style_SFoyer              0.0783      0.021      3.786      0.000       0.038       0.119
House_Style_Two_Story           0.2276      0.012     18.259      0.000       0.203       0.252
House_Style_Two_and_Half_Fin    0.3505      0.059      5.916      0.000       0.234       0.467
Kitchen_AbvGr                  -0.1559      0.017     -8.933      0.000      -0.190      -0.122
Latitude                        0.5065      0.191      2.652      0.008       0.132       0.881
Longitude                      -0.3757      0.139     -2.702      0.007      -0.648      -0.103
Lot_Area                     7.673e-06   8.81e-07      8.714      0.000    5.95e-06     9.4e-06
Lot_Frontage                    0.0003      0.000      3.031      0.002       0.000       0.001
Overall_Cond_Below_Average     -0.1412      0.018     -7.790      0.000      -0.177      -0.106
Overall_Cond_Excellent          0.2731      0.026     10.575      0.000       0.222       0.324
Overall_Cond_Fair              -0.3190      0.028    -11.362      0.000      -0.374      -0.264
Overall_Cond_Good               0.1045      0.010     10.534      0.000       0.085       0.124
Overall_Cond_Poor              -0.3461      0.063     -5.494      0.000      -0.470      -0.223
Overall_Cond_Very_Good          0.1370      0.014      9.502      0.000       0.109       0.165
Overall_Cond_Very_Poor         -0.2459      0.100     -2.460      0.014      -0.442      -0.050
Total_Bsmt_SF                   0.0002   1.37e-05     15.900      0.000       0.000       0.000
Year_Built                      0.0039      0.000     21.225      0.000       0.004       0.004
==============================================================================
Omnibus:                       53.628   Durbin-Watson:                   2.006
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               96.843
Skew:                           0.209   Prob(JB):                     9.35e-22
Kurtosis:                       4.015   Cond. No.                     5.58e+07
```

Fig. 11. Summary of Model 3

positive value, we can say that if we increase the value of fireplaces by one unit, the predicted value of the dependent variable, i.e., sale Price will increase by 0.0755, keeping all the other independent variables constant.

The p-values of all the independent variables in this model are less than 0.05 which means these variables are significant for predicting the value of dependent variable y. If any variable has a high p-value i.e. ¿ 0.05 it means that there is some non-zero co-relation with the dependent variable y.

VII. Diagnostics

In the context of Gauss-Markov's Linear Regression, we consider four key assumptions:

1) **Linearity:** There should be a linear relationship between the dependent and independent variables.
2) **Multicollinearity:** There should not be any correlation between independent variables.
3) **Homoscedasticity:** The variance of the errors should be constant over all independent variables.
4) **Zero Mean Condition:** The average of all the errors should be equal to zero.

In the above-mentioned assumptions, I checked the multi-collinearity with the help of the variance inflation factor (VIF) and found that our important features were not dependent on each other. Hence there is no multicollinearity.

I have also checked the correlation between Sale_Price_Actual and Sale_Price_Predicted as we can see on the training set and we found that there is a linear relationship between actual and predicted values as we can see in Fig. 12.

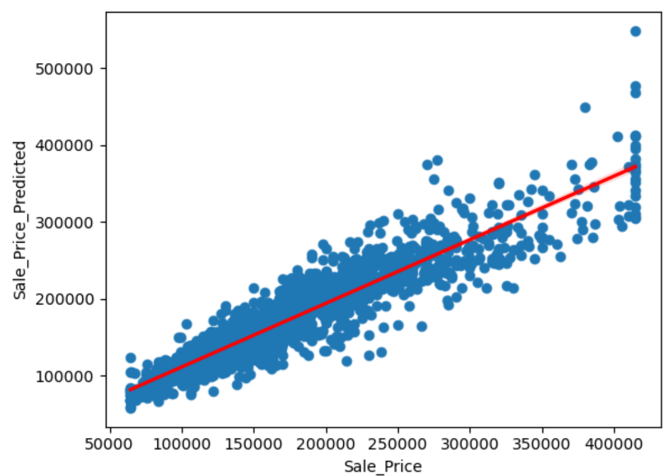


Fig. 12. Correlation between Actual and Predicted (Training Data)

VIII. Evaluation

I have also checked the correlation between Sale_Price_Actual and Sale_Price_Predicted as we can see on the test set, we found that there is a linear relationship between actual and predicted values, as we can see in Fig. 13.

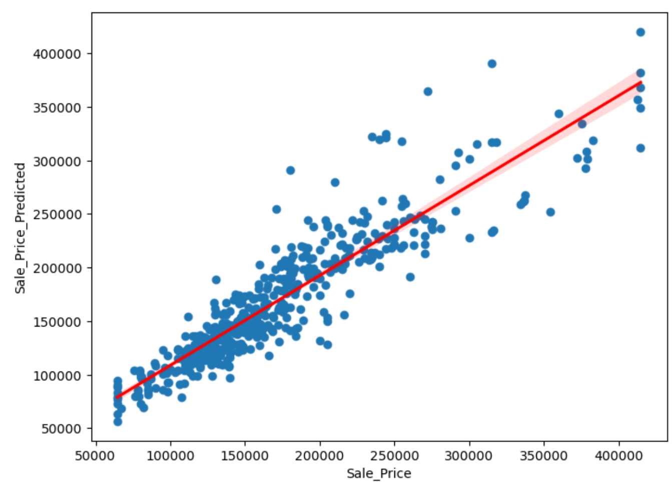


Fig. 13. Correlation between Actual and Predicted (Test Data)

IX. Evaluation

Finally, I have evaluated the accuracy of the model, and the following accuracy measures were considered:

- **MAE (Mean of Absolute Errors)**
- **MAPE (Mean Absolute Percentage of Errors)**
- **MSE (Mean Squared Errors)**
- **RMSE (Root Mean Squared Errors)**
- **R2 Score (Goodness of Fit)**

The calculated values for both the training and test sets are as follows:

Train MSE $= 706052127.63$ | Test MSE $= 724446318.28$

Train MAE $= 18923.75$ | Test MAE $= 18756.92$

Train MAPE $= 0.10$ | Test MAPE $= 0.10$

Train RMSE $= 26571.64$ | Test RMSE $= 26915.54$

Train R2_Score $= 0.84$ | Test R2_Score $= 0.83$

In the above output, an R2 score of 0.84 on the training data and 0.83 on the test data indicates that our final model, i.e., Model 3, is providing good accuracy of 84% and 83% on the training and test sets, respectively.