Investigating factors influencing the housing prices using multiple linear regression

Sushma Suresh Hebbalakar 22224122

***Abstract*— Housing price prediction is a case of exploratory data analysis. Here we are considering housing prices in the given data set and using variety of linear models to predict and analyze the trend of housing prices. I have implemented a distributed housing price prediction analysis model.Through the comparison of accuracy of multiple linear models, found out the best linear analysis model with the best prediction results**

1 INTRODUCTION

Housing price prediction is a case of exploratory data analysis. Here we are considering housing prices in the given data set, and using variety of linear models to predict and analyze the trend of housing prices.

Housing price prediction has become very essential thing in recent days.There are many factors that can affect the price of a house, such as location, size, age, condition, and many more.Predicting the price of a house can be done using different methods.The models can help to identify the significant determinants that affect housing prices and can be used to determine the optimal price range of a property which are useful for decision-making for both buyers and sellers.

OLS stands for Ordinary Least Squares, and it is a method used in statistics for estimating the parameters of a linear regression model. In the context of linear regression, Ordinary Least Squares refers to the method of finding the line that minimizes the sum of the squared differences between the observed and predicted values.The "ordinary" in OLS implies that it is a standard and widely used approach.The resulting equation represents the "best-fitting" line through the data. The linear regression equation in a simple case with one independent variable (X) and one dependent variable (Y) is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon \qquad (1)$$

where:

- Y is the dependent variable.
- X is the independent variable.
- β_0 is the intercept (constant term).
- β_1 is the slope coefficient.
- ε is the error term.

2 METHODOLOGY

2.1 Overview

The CRISP-DM [4] framework is generally used in data analysis projects and it has a number of phases which are iterative. It is not uncommon to redo a phase.

Business Understanding:: The business problem is identified and it is transformed into a data analysis problem.

Data Understanding:: The quality of the data available must be assessed and problems or changes needed are identified.

Data Preparation:: Data must be useable in the modelling phase. Tasks at this point usually include variable selection, transformation, dealing with missing values or feature engineering.

Modelling:: Modelling: The data is modelled at this stage and the output is reviewed. This will include checking that the key modelling assumptions have been met and how the model is performing.

Deployment:: When the final model is agreed it can then be deployed for use in the business.Except for deployment all phases of the CRISP – DM framework have been followed.

A. 2.2 Business Understanding

The business problem is to predict the Sale_Price of the house. Knowing the key variables will help estimate the Sale_Price. In terms of a data analysis problem a dependent variable must be estimated using several independent variables, this is the general purpose of the MLR modelling technique.

2.3 Data Understanding

The first stage in the analysis is to understand the available data that has been made available by the card issuer. There are 2413 records 18 variables in the given data set.The variables in the data set are as below Table 1:

Variable/Features	Description	Data Type
Lot_Frontage	The amount of the street connected to property in feet	int64
Lot_Area	The size of the plot in square feet	int64
Bldg_Type	The type of dwelling	object
House_Style	The style of dwelling	object
Overall_Cond	The overall condition of the house	object
Year_Built	The year the house was constructed	int64
Exter_Cond	The condition of the material on the exterior of the house	object
Total_Bsmt_SF	Total basement area in square feet	int64
First_Flr_SF	Ground floor area in square feet	int64
Second_Flr_SF	First floor area in square feet	int64
Full_Bath	Number of full bathrooms	int64
Half_Bath	Number of half bathrooms	int64
Bedroom_AbvGr	Number of bedrooms on or above ground floor	int64
Kitchen_AbvGr	Number of kitches on or above ground floor	int64
Fireplaces	Number of fireplaces	int64
Longitude	Longitude of plot	float64
Latitude	Latitude of plot	float64
Sale_Price	The sale price of the house	int64

Table 1: Data description

To understand the data the following descriptive statistics were calculated.

Column1	Lot_Frontage	Lot_Area	Year_Built	Total_Bsmt_SF	First_Flr_SF	Second_Flr_SF	Full_Bath
count	2413	2413	2413	2413	2413	2413	2413
mean	55.462495	10060.208	1969.43639	1022.8276	1133.857024	339.242851	1.539163
std	33.542416	8222.7599	29.487943	408.977848	366.44247	423.198488	0.544667
min	0	1300	1872	0	334	0	0
25%	3700%	739000%	195300%	78400%	86600%	0%	100%
50%	6000%	936000%	197100%	97000%	106000%	0%	200%
75%	7700%	1140400%	199800%	124600%	135000%	70400%	200%
max	313	215245	2010	3206	3820	1872	4

Table 2: Descriptive statistics for numerical values

Column1	Half_Bath	Bedroom_A	Kitchen_Abv	Fireplaces	Longitude	Latitude	Sale_Price
count	2413	2413	2413	2413	2413	2413	2413
mean	0.377953	2.854538	1.040199	0.602984	-93.642391	42.03373	175567.64
std	0.498467	0.81343	0.200642	0.648911	0.026159	0.018015	70979.614
min	0	0	0	0	-93.693153	41.986498	35000
25%	0%	200%	100%	0%	-9366%	4202%	12950000%
50%	0%	300%	100%	100%	-9364%	4203%	15900000%
75%	100%	300%	100%	100%	-9362%	4205%	20690000%
max	2	6	3	4	-93.577427	42.063381	755000

Table 2: Descriptive statistics for numerical values(continued)

Variables 'Lot_Frontage' appear to be normally distributed based on the skew .As variables 'Full_Bath', 'Half_Bath', 'Bedroom_AbvGr', 'Kitchen_AbvGr', 'Fireplaces' are ordinal variables normal distribution does not apply. Other variables are right skewed or left skewed see Table 3.

Skewness of Lot_Frontage	-0.081064024
Skewness of Lot_Area	13.38559308
Skewness of Year_Built	-0.586359311
Skewness of Total_Bsmt_SF	0.456102043
Skewness of First_Flr_SF	1.042652395
Skewness of Second_Flr_SF	0.803617849
Skewness of Full_Bath	0.24486056
Skewness of Half_Bath	0.663579564
Skewness of Bedroom_AbvGr	0.184044239
Skewness of Kitchen_AbvGr	4.67866351
Skewness of Fireplaces	0.738615919
Skewness of Longitude	-0.337302874
Skewness of Latitude	-0.507462408
Skewness of Sale_Price	1.744272889

Table 3: Visualizing skewness with respect to Sale_Price

Checking the counts/frequencies for nominal and categorical value Figure1:

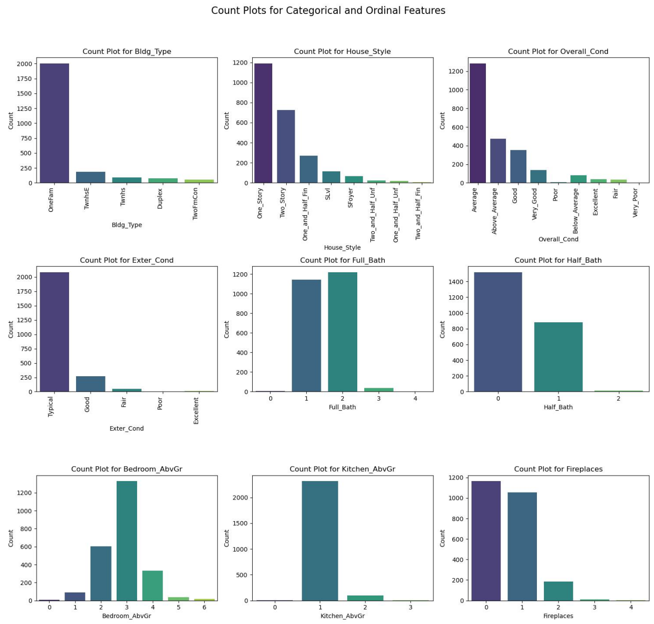

Figure 1: Counts for categorical and nominal

Checking the correlation of numerical features with each

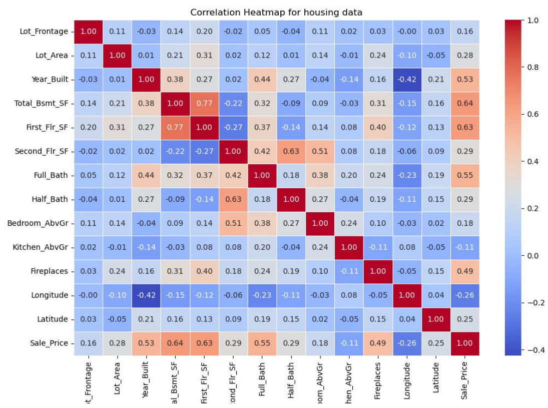

Figure 2: Correlation of numerical features

Visualizing the correlation of all numeric features with respect to Sale_Price Figure3 : Numeric Features Correlation with Sale_Price

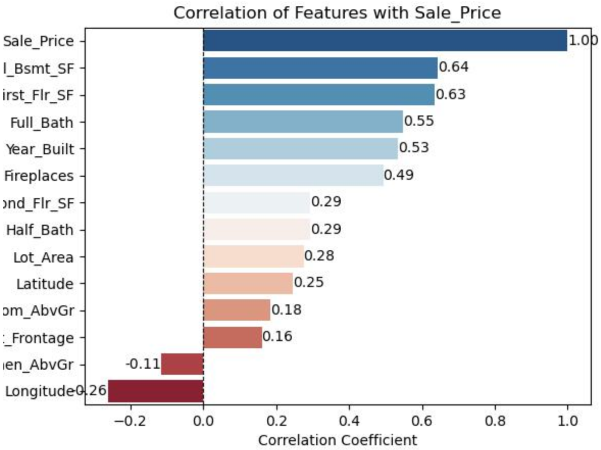

Figure 3: Correlation matrix with respect to

Visualizing distribution of all features using scatter plot for numerical features and box plot for categorical features:

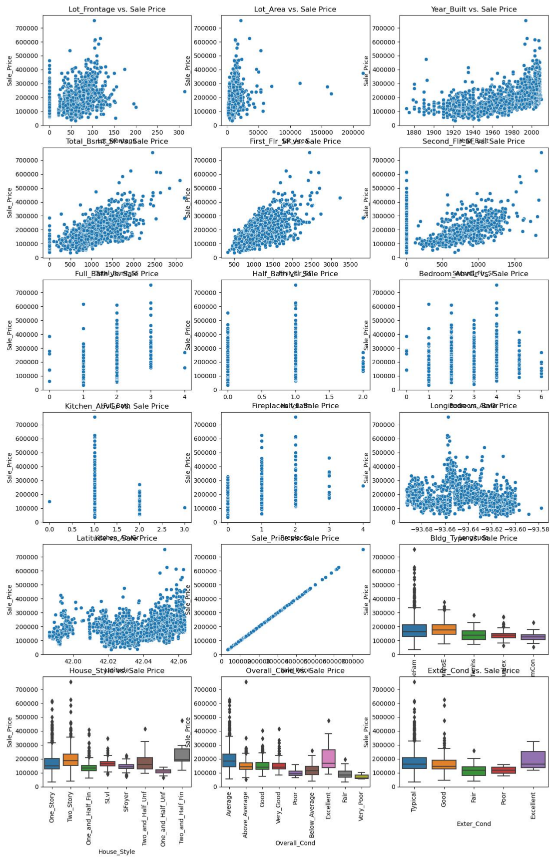

Figure 4: Visualizing data distribution

Visualizing the data distribution of all numerical features using histogram

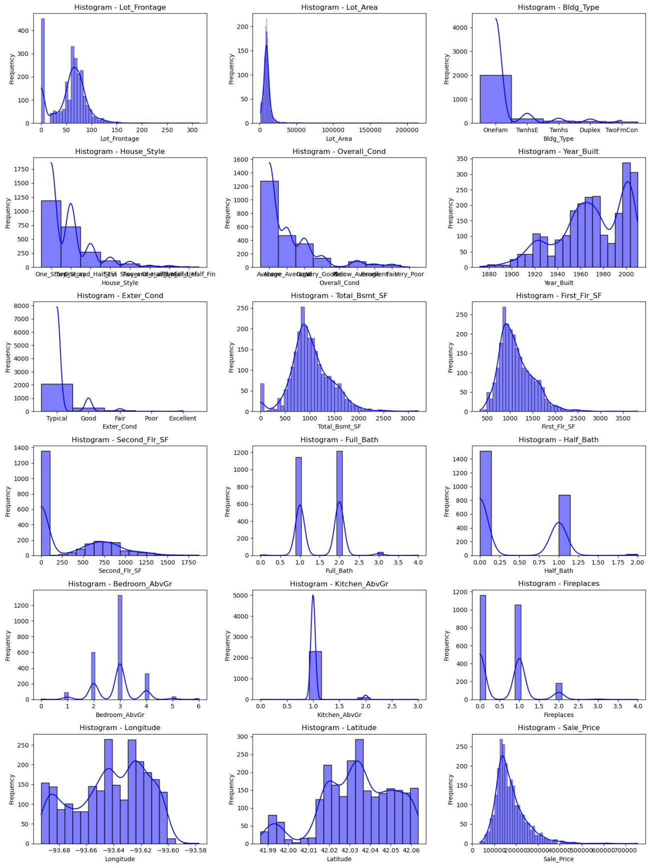

Figure 5: Histplot numrical features

B. 2.4 Data Preparation

With understanding of the data, it is then prepared for the modelling phase. Inspection of the data showed that the following needs to be done:

- Encoding of categorical variables Bldg_Type, House_Style, Overall_Cond, Exter_Cond to be done.
- Based on skewness and type of data of variables First_Flr_SF, Second_Flr_SF, Sale_Price is required to normalize the data.

I am not doing pre-processing all together instead I'll go step by step. First I am encoding the categorical data using Label encoder below table 4 shows how data has been mapped.

	Mapping for Bldg_Type	House_Style	Overall_Cond	Exter_Cond
0	Duplex	One_Story'	Above_Average	Excellent
1	OneFam	One_and_Half_Fin	Average	Fair
2	Twnhs	One_and_Half_Unf	Below_Average	Good
3	TwnhsE	SFoyer	Excellent	Poor
4	TwoFmCon	SLvl	Fair	Typical
5		Two_Story	Good	
6		Two_and_Half_Fin	Poor	
7		Two_and_Half_Unf	Very_Good	
8			Very_Poor	

Table 4: The mapping values after encoding

2.5 Modelling

To model the data multiple linear regression is being used and a common approach is to build the best linear unbiased model using the ordinary least squares basis. The basis revolves around the Gauss – Markov assumptions which seek a MLR model that:

- Is in the correct functional form;
- Has errors with a constant variance i.e., homoscedasticity and
- Has no autocorrelation between the errors.

There are a few other key additional assumptions when building a MLR which are:

- The errors are normally distributed;
- There is no multicollinearity between the predictor (independent) variables
- There are no influential data points in the data.

The first assumption "correct functional form" has been tested and proven, see figure 6, by showing that there is correlation between the dependent variable and one or more independent variable. The remaining assumptions will be tested as follows:

Condition	Test	Pass Value
Homoscedasticity	Plot standard re sidual value to standard pre dicted value	Randomness should be shown with most values in or around 0.
No autocorrelation between errors	Durbin Watson test	Approx. 2 is best value. 3 is concerning.
Errors are normally distributed	Run P-P plots or histogram	Errors should show a normal distribution.
No multicollinearity between the predictor (independent) variables	Run VIF test	A max of approx. 2 is best with val ues >= 5 being a problem.

Table 5: Multiple Linear Regression Assumptions.

During the model build process the model output will be evaluated against the assumptions summarised in table 5 and where one test fails the model will be updated so that the final model produced is appropriate and meets the required criteria. The best model will be chosen using the following criteria:

- Gauss Markov assumptions are met;
- Other key assumptions are met;
- Highest adjusted R2 value; and
- Lowest Standard Error.

2.6 Evaluation

The final model produced will be compared to the business problem to ensure that the model prepared is sensible in the context of the business problem. This will be done by analysing the coefficients of the linear model and the variables.

3 MODEL OUTPUT AND DISCUSSION

3.1 Model 1

The first model built used the independent variables correlated to the dependent variable, as shown in Figure 3.The original dataset values for these variables were entered into the model. The output is as follows:

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.12e+07	3.61e+06	-3.101	0.002	-1.83e+07	-4.12e+06
Lot_Frontage	119.8863	21.780	5.504	0.000	77.170	162.603
Lot_Area	0.5335	0.099	5.412	0.000	0.340	0.727
Bldg_Type	-1530.5464	1038.001	-1.475	0.141	-3566.366	505.273
House_Style	724.0518	562.028	1.288	0.198	-378.248	1826.352
Overall_Cond	2947.2095	376.620	7.825	0.000	2208.549	3685.870
Year_Built	676.9405	37.466	18.068	0.000	603.458	750.423
Exter_Cond	-1123.1065	936.256	-1.200	0.230	-2959.376	713.163
Total_Bsmt_SF	46.1508	2.723	16.951	0.000	40.811	51.491
First_Flr_SF	92.7108	3.565	26.002	0.000	85.718	99.704
Second_Flr_SF	92.4305	3.842	24.055	0.000	84.894	99.967
Full_Bath	199.1959	2042.735	0.098	0.922	-3807.199	4205.591
Half_Bath	-1283.6969	2120.672	-0.605	0.545	-5442.949	2875.555
Bedroom_AbvGr	-1.629e+04	1139.687	-14.294	0.000	-1.85e+04	-1.41e+04
Kitchen_AbvGr	-3.518e+04	3894.155	-9.034	0.000	-4.28e+04	-2.75e+04
Fireplaces	6848.0435	1296.151	5.283	0.000	4305.917	9390.170
Longitude	-6.172e+04	3.07e+04	-2.008	0.045	-1.22e+05	-1431.380
Latitude	9.861e+04	4.24e+04	2.326	0.020	1.55e+04	1.82e+05

Figure 6: Model 1 Summary

Below are some other statistical observations:

- Mean Squared Error: 869059497.6025758
- R-squared: 0.8379090793353802
- Mean Absolute Error: 20495.726954279242
- F-statistic:488.6
- Durbin-Watson:2.019
- Kurtosis:8.384

- Overall Model 1 performed well based on the adjusted R2 basis since it explained 0.82 (61%) of the variation.
- The VIF is high for several variables which suggests that some multicollinearity is present.
- The Durbin Watson test suggests that there is no auto correlation between errors.
- Since the Mean Squared Error and Mean Absolute Error is high the model is not significant

The plot in figure 7 to check for homoscedasticity that the errors (residuals) are randomly distributed or not and it depicts they are not randomly distributed.In PP-Plot deviations from a straight line may indicate departures from normality.In Figure 8 the points deviate upward or downward, it might suggest skewness or heavy tails.Model 1 failed the homoscedasticity and it is deemed unsuccessful so a new model was prepared.

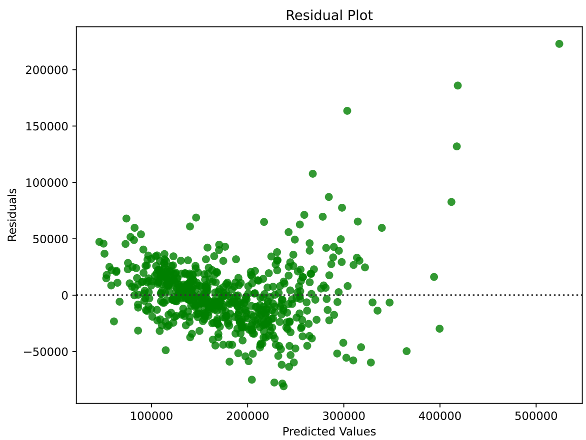

Figure 7: Homoscedasticity for Model1

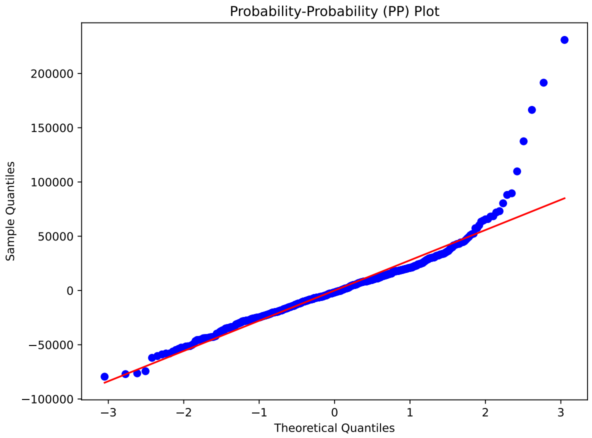

Figure 8: PP-Plot for Model1

C. 3.2 Model 2

To address the homoscedasticity issue which, is likely due to the skewness of the original variables log transformation has been applied to 'First_Flr_SF', 'Second_Flr_SF', 'Sale_Price' as these had positive skewness and are not ordinal values.Below figures 9 and 10 depicts the distribution of data before transformation and after transformation.

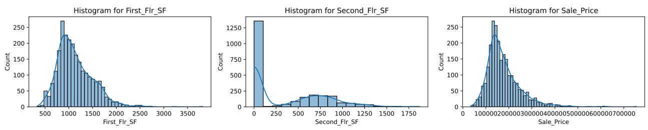

Figure 9: Skewness before tranformation

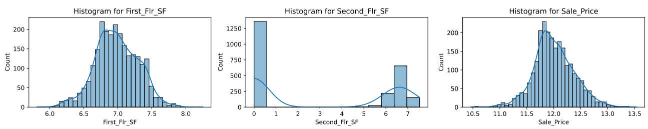

Figure 10: Skewness after tranformation

The logN of 'First_Flr_SF', 'Second_Flr_SF', 'Sale_Price' variables.All the independent variables as per model 1 were used. The output of model 2 is depicted in Table7 :

	coef	std err	t	P>|t|	[0.025	0.975]
const	-38.0961	18.970	-2.008	0.045	-75.302	-0.890
Lot_Frontage	0.0003	0.000	2.960	0.003	0.000	0.001
Lot_Area	0.0913	0.009	9.733	0.000	0.073	0.110
Bldg_Type	0.0130	0.006	2.242	0.025	0.002	0.024
House_Style	0.0163	0.003	5.735	0.000	0.011	0.022
Overall_Cond	0.0192	0.002	9.760	0.000	0.015	0.023
Year_Built	0.0045	0.000	23.159	0.000	0.004	0.005
Exter_Cond	0.0043	0.005	0.873	0.383	-0.005	0.014
Total_Bsmt_SF	0.0445	0.003	13.841	0.000	0.038	0.051
First_Flr_SF	0.6352	0.020	31.773	0.000	0.596	0.674
Second_Flr_SF	0.0353	0.002	14.821	0.000	0.031	0.040
Full_Bath	0.0666	0.010	6.416	0.000	0.046	0.087
Half_Bath	0.0382	0.011	3.547	0.000	0.017	0.059
Bedroom_AbvGr	-0.0482	0.006	-8.107	0.000	-0.060	-0.037
Kitchen_AbvGr	-0.1677	0.021	-8.122	0.000	-0.208	-0.127
Fireplaces	0.0592	0.007	8.750	0.000	0.046	0.072
Longitude	-0.2040	0.162	-1.258	0.208	-0.522	0.114
Latitude	0.3905	0.221	1.768	0.077	-0.043	0.824

Table 7: Model 2 Summary

Below are some other statistical observations:

- Mean Squared Error: 0.021116829192019192
- R-squared: 0.8471160826168324
- Mean Absolute Error: 0.10939595781509125
- F-statistic:499.5
- Durbin-Watson:2.002
- Kurtosis:5.599

Model 2 has performed better than model 1. There is a 0.01 increase in the adjusted R2 and the standard error of the estimate in model 2 has reduced from 869059497.6025758 to 0.021.

Below are actual versus predicted value plot, scatter plot for homoscedasticity and PP-Plot

Figure 11: Actual versus Predicted plot

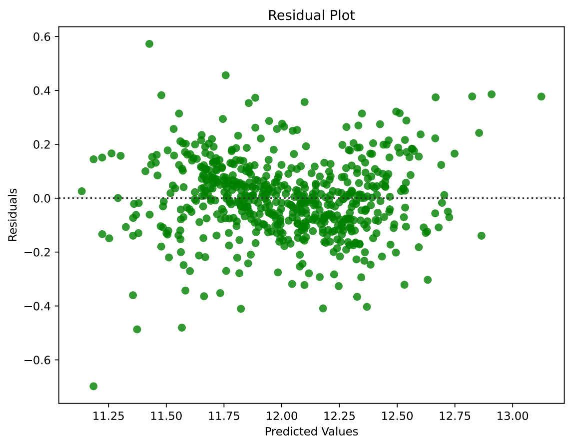

Figure 12: Homoscedasticity for Model2

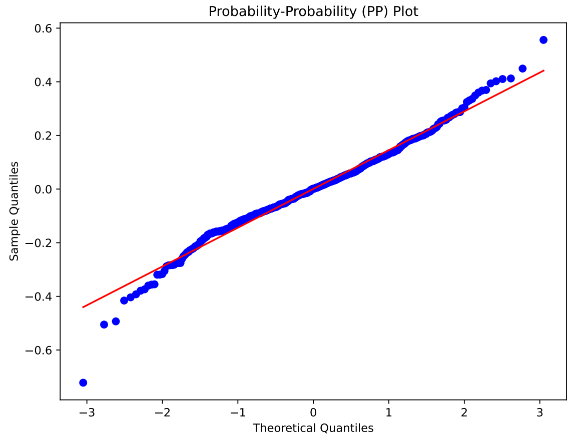

Figure 13: PP-Plot for Model2

In figure 9 the homoscedasticity assumption appears to be met as the errors look to be randomly distributed. The curve in the P-P plot(Figure 10) for the residuals is close to the cumulative plot line though there is some variance present.**A note of concern is that there are few features which have p-value greater than 0.05.** So for our next model we are eleminating the features that have p-value greater than 0.05

D. 3.3 Model 3

The 3rd model is built on all the features accept the one which had greater than 0.05 p-value in model2. For model3 the output is as follows:

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.8369	0.382	-7.432	0.000	-3.586	-2.088
Lot_Frontage	0.0003	0.000	2.952	0.003	0.000	0.001
Lot_Area	0.0919	0.009	9.971	0.000	0.074	0.110
Bldg_Type	0.0129	0.006	2.226	0.026	0.002	0.024
House_Style	0.0162	0.003	5.697	0.000	0.011	0.022
Overall_Cond	0.0190	0.002	9.759	0.000	0.015	0.023
Year_Built	0.0047	0.000	25.814	0.000	0.004	0.005
Total_Bsmt_SF	0.0448	0.003	13.932	0.000	0.038	0.051
First_Flr_SF	0.6368	0.020	31.983	0.000	0.598	0.676
Second_Flr_SF	0.0354	0.002	14.878	0.000	0.031	0.040
Full_Bath	0.0677	0.010	6.567	0.000	0.047	0.088
Half_Bath	0.0398	0.011	3.709	0.000	0.019	0.061
Bedroom_AbvGr	-0.0486	0.006	-8.186	0.000	-0.060	-0.037
Kitchen_AbvGr	-0.1682	0.021	-8.152	0.000	-0.209	-0.128
Fireplaces	0.0594	0.007	8.801	0.000	0.046	0.073

Table 8: Model 3 summary

Below are some other statistical observations:

- Mean Squared Error: 0.02129925625039484
- R-squared: 0.8457953273525098
- Mean Absolute Error: 0.1100927998734349
- F-statistic:605.5
- Durbin-Watson:605.5
- Kurtosis:5.619

Below are actual versus predicted value plot, scatter plot for homoscedasticity and PP-Plot

Figure 14: Actual versus Predicted plot

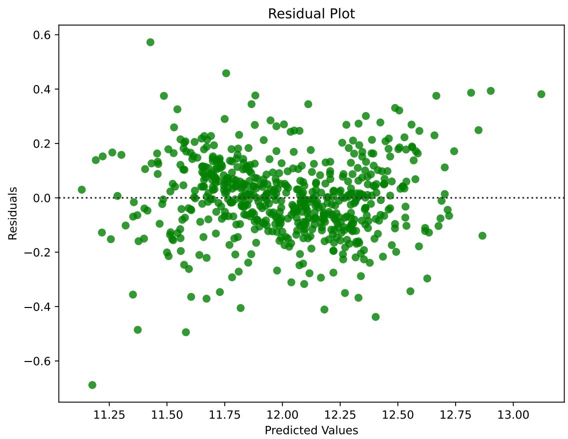

Figure 15: Homoscedasticity for Model3

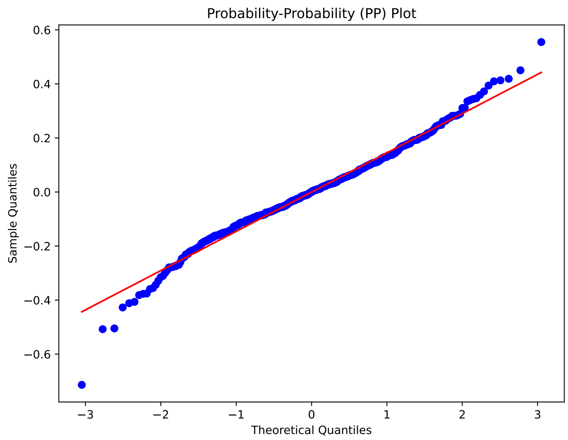

Figure 16: PP-Plot for Model3

In figure 15 the homoscedasticity assumption appears to be met as the errors look to be randomly distributed. The curve in the P-P plot(Figure 16) for the residuals is close to the cumulative plot.

Though model 3 has performed well and has passed the diagnostic tests as described in table 5. But, there is increase in mse by 0.01 and decrease in R2 by 0.02, but also the f-static has increased from 499 to 605 when compared to model2.

E. 3.4 Model 4

In the 4th model based on observation with respect to correlation.As 'Bldg_Type' had less correlation I have tried removing that feature and train the model.Below are the outputs obtained:

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.7756	0.381	-7.282	0.000	-3.523	-2.028
Lot_Frontage	0.0003	0.000	2.783	0.005	9.44e-05	0.001
Lot_Area	0.0844	0.009	9.827	0.000	0.068	0.101
House_Style	0.0163	0.003	5.729	0.000	0.011	0.022
Overall_Cond	0.0187	0.002	9.633	0.000	0.015	0.023
Year_Built	0.0047	0.000	25.787	0.000	0.004	0.005
Total_Bsmt_SF	0.0452	0.003	14.058	0.000	0.039	0.051
First_Flr_SF	0.6405	0.020	32.250	0.000	0.602	0.679
Second_Flr_SF	0.0357	0.002	15.028	0.000	0.031	0.040
Full_Bath	0.0691	0.010	6.717	0.000	0.049	0.089
Half_Bath	0.0388	0.011	3.620	0.000	0.018	0.060
Bedroom_AbvGr	-0.0515	0.006	-8.879	0.000	-0.063	-0.040
Kitchen_AbvGr	-0.1685	0.021	-8.160	0.000	-0.209	-0.128
Fireplaces	0.0599	0.007	8.864	0.000	0.047	0.073

Table 9: Model4 summary

Other statistical values:

- Mean Squared Error: 0.02114616034836353
- R-squared: 0.8469037277200566
- Mean Absolute Error: 0.10976159296680633
- F-statistic:650.3
- Durbin-Watson:2.016
- Kurtosis:5.627

The mse has decreased by 0.001 and R2 has increased by 0.001 along with increment in F-static value which means model4 has performed better than model3

Below are actual versus predicted value plot, scatterplot for homoscedasticity and PP-Plot

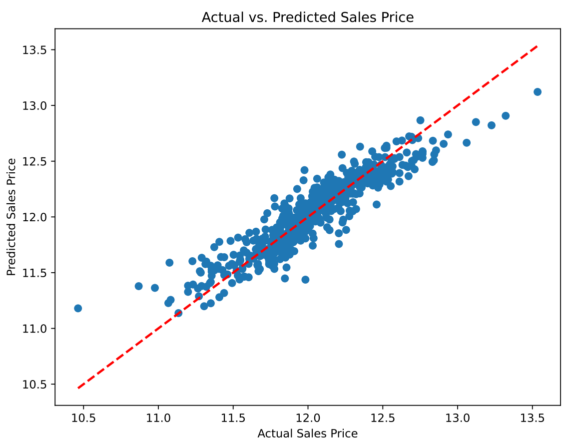

Figure 17: Actual versus Predicted plot

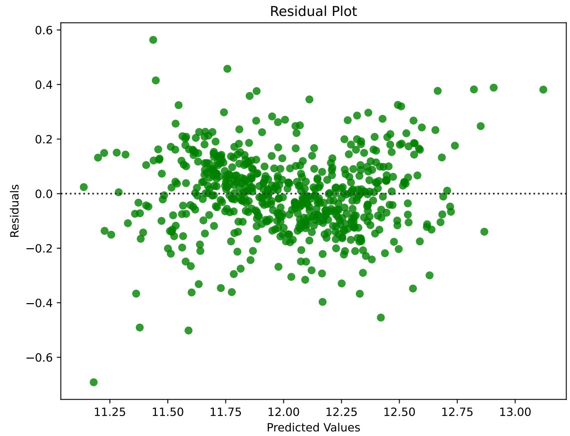

Figure 18: Homoscedasticity

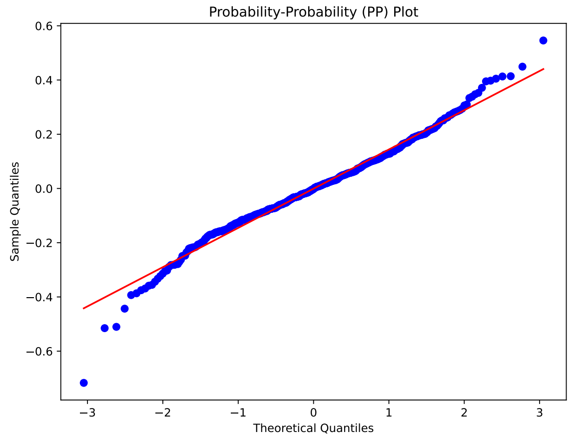

Figure 19: PP-Plot.

3.5 Model5

In the 5th model again based on observation with respect correlation I am trying to remove Overall_Cond featue.Below are the outputs obtained:

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.7006	0.374	-4.551	0.000	-2.433	-0.968
Lot_Frontage	0.0003	0.000	2.591	0.010	7.42e-05	0.001
Lot_Area	0.0871	0.009	9.894	0.000	0.070	0.104
House_Style	0.0164	0.003	5.615	0.000	0.011	0.022
Year_Built	0.0042	0.000	23.420	0.000	0.004	0.005
Total_Bsmt_SF	0.0451	0.003	13.689	0.000	0.039	0.052
First_Flr_SF	0.6375	0.020	31.311	0.000	0.598	0.677
Second_Flr_SF	0.0362	0.002	14.865	0.000	0.031	0.041
Full_Bath	0.0702	0.011	6.653	0.000	0.050	0.091
Half_Bath	0.0366	0.011	3.334	0.001	0.015	0.058
Bedroom_AbvGr	-0.0524	0.006	-8.808	0.000	-0.064	-0.041
Kitchen_AbvGr	-0.1870	0.021	-8.872	0.000	-0.228	-0.146
Fireplaces	0.0578	0.007	8.344	0.000	0.044	0.071

Table10:Model 5 Summary

Other statistical values:

- Mean Squared Error: 0.021686266851140678
- R-squared: 0.8429934059005336
- Mean Absolute Error: 0.11229916263357065
- F-statistic:662.9
- Durbin-Watson:2.021
- Kurtosis:5.048

Below are actual versus predicted value plot, scatter plot for homoscedasticity and PP-Plot

Figure 20: Actual versus Predicted plot

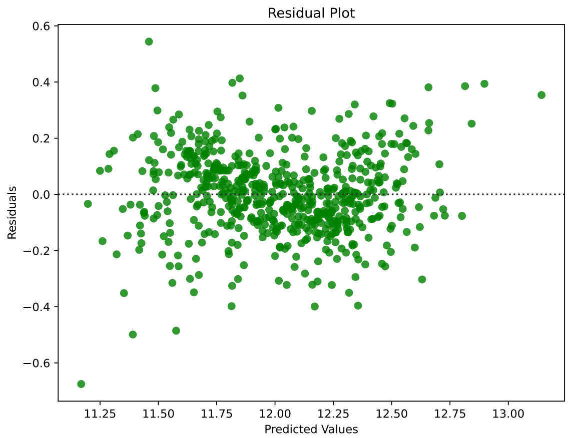

Figure 21: Homoscedasticity

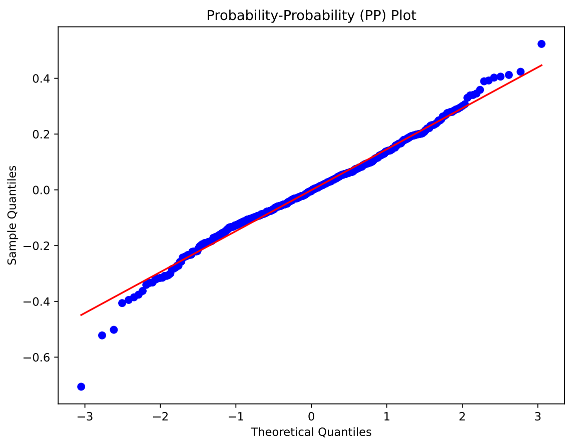

Figure 22: PP-plot

The mse has increased by 0.005 and R2 has decreased by 0.004 along with increment in F-static value which means model5 has performed better but I would consider model4 as final model.

4 Evaluation

The purpose of this analysis was to define a model that could be used to estimate the Sale_Price of the house for given features. The best model produced that met the validation criteria has the following formula:

$\log(NSale\_Price) = 0.0003 \cdot \text{(Lot\_Frontage)} + 0.0844 \cdot \text{(Lot\_Area)} + 0.0163 \cdot \text{(House\_Style)} + 0.0187 \cdot \text{(Overall\_Cond)} + 0.0047 \cdot \text{(Year\_Built)} + 0.0452 \cdot \text{(Total\_Bsmt\_SF)} + 0.6405 \cdot \log(\text{ N First\_Flr\_SF}) + 0.0357 \cdot \log(\text{N Second\_Flr\_SF}) + 0.0691 \cdot \text{(Full\_Bath)} + 0.0388 \cdot \text{(Half\_Bath)} - 0.0515 \cdot \text{(Bedroom\_AbvGr)} - 0.1685 \cdot \text{(Kitchen\_AbvGr)} + 0.0599 \cdot \text{(Fireplaces)} + 0.0211$

5 CONCLUSIONS

The business problem posed was to develop a method to predict Sale_Price of house. Using regression modelling where the Sale_Price was modelled based on independent variables. R^2 is approximately 0.8469, suggesting that the model explains about 84.69% of the variability in the data. MSE is a measure of the average squared difference between the actual and predicted values. It quantifies the average magnitude of errors in Model3 this value is 0.021. MAE is approximately 0.1098, indicating, on average, the absolute difference between the actual and predicted values is about 0.1098.

Over all Model4 with Lower values for MSE and MAE are desirable, indicating better performance. Higher R^2 values indicate better explanatory power of the model.

References

[1] S. Sharma, D. Arora, G. Shankar, P. Sharma and V. Motwani, "House Price Prediction using Machine Learning Algorithm," 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2023, pp. 982-986, doi: 10.1109/ICCMC56507.2023.10084197.

[2] X. Yang, Z. Yin and J. Li, "Housing Price Mathematical Prediction Method through Big Data Analysis and Improved Linear Regression Model," 2021 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), Shenyang, China, 2021, pp. 751-754, doi: 10.1109/TOCS53301.2021.9688839.

[3] T. Alshammari, "Evaluating machine learning algorithms for predicting house prices in Saudi Arabia," 2023 International Conference on Smart Computing and Application (ICSCA), Hail, Saudi Arabia, 2023,pp. 1-5, doi: 10.1109/ICSCA57840.2023.10087486.