House Sale Price Prediction

MSc in Data Analytics January 2021

Sanket Sonu

x19206071@student.ncirl.ie

Abstract: **This report shows the relationship between house characteristics and sales price which is predicted by building model with Multiple Regression technique.**

I. OBJECTIVES:

- Use descriptive statistics and appropriate visualisations to enhance understanding of the variables in the dataset.
- Describe the model building steps you undertook in the process of arriving at your final regression model. The rationale for rejecting intermediate models should be explained clearly and details provided on treatment of outliers, transformations undertaken etc.
- Provide details on diagnostics undertaken to verify that the Gauss Markov and other relevant assumptions of multiple regression have been satisfied.
- Provide a succinct summary of the parameters of your final model and details of model performance and fit.

II. DESCRIPTION OF DATASET:

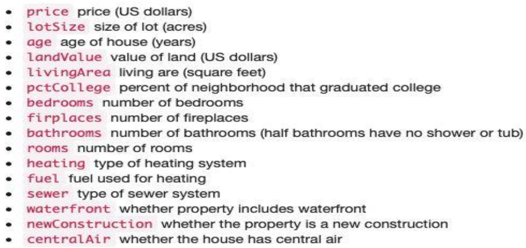
The given dataset of House Details, which includes various features like price, lotSize, age, landValue, livingArea, pctCollege, bedrooms, fireplaces, bathrooms, rooms, heatingfuel, sewer, waterfront, newConstruction, centralAir.

Number of columns: 16

Number of observations: 1728

Dependent Variable: Price

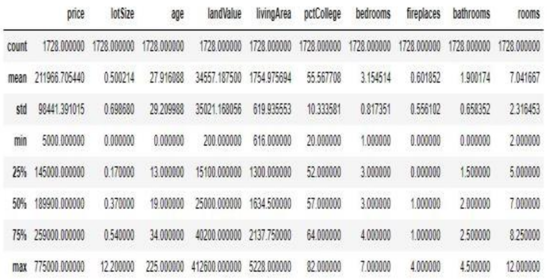
Below screenshot is Variables details:



- `price` price (US dollars)
- `lotSize` size of lot (acres)
- `age` age of house (years)
- `landValue` value of land (US dollars)
- `livingArea` living are (square feet)
- `pctCollege` percent of neighborhood that graduated college
- `bedrooms` number of bedrooms
- `firplaces` number of fireplaces
- `bathrooms` number of bathrooms (half bathrooms have no shower or tub)
- `rooms` number of rooms
- `heating` type of heating system
- `fuel` fuel used for heating
- `sewer` type of sewer system
- `waterfront` whether property includes waterfront
- `newConstruction` whether the property is a new construction
- `centralAir` whether the house has central air

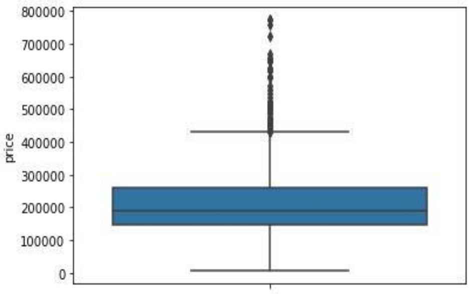
III. DATA VISUALIZATION AND DISCRIPTIVE STATISTICS:

From the below fig, it is clearly observed that the total count of data is 1728. The mean is 211966.705 and the Standard Deviation is 98441.391. Min, Max and IQR values are also displayed.



	price	lotSize	age	landValue	livingArea	pctCollege	bedrooms	fireplaces	bathrooms	rooms
count	1728.000000	1728.000000	1728.000000	1728.000000	1728.000000	1728.000000	1728.000000	1728.000000	1728.000000	1728.000000
mean	211966.705440	0.500214	27.916088	34557.187500	1754.975694	55.567708	3.154514	0.601852	1.900174	7.041667
std	98441.391015	0.698680	29.209988	35021.168056	619.935553	10.333581	0.817351	0.556102	0.658352	2.316453
min	5000.000000	0.000000	0.000000	200.000000	616.000000	20.000000	1.000000	0.000000	0.000000	2.000000
25%	145000.000000	0.170000	13.000000	15100.000000	1300.000000	52.000000	3.000000	0.000000	1.500000	5.000000
50%	189900.000000	0.370000	19.000000	25000.000000	1634.500000	57.000000	3.000000	1.000000	2.000000	7.000000
75%	259000.000000	0.540000	34.000000	40200.000000	2137.750000	64.000000	4.000000	1.000000	2.500000	8.250000
max	775000.000000	12.200000	225.000000	412600.000000	5228.000000	82.000000	7.000000	4.000000	4.500000	12.000000

Boxplot is used to graphically show the numerical data through their quartiles. This also demonstrates, how tightly data is grouped. Outliers are also observed in this boxplot.



To Calculate IQR of Price, we need values of Q1 and Q3 of Price variable:

Q1 – 145000

Q3 – 259000

IQR = Q3 – Q1 = 114000

Correlation value ranges from -1 to 1, where

1 = Positive correlation

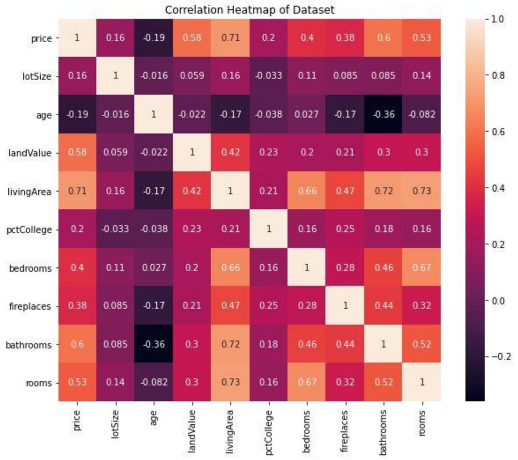
0 = No correlation

-1 = Negative correlation

Heatmap show the correlation between variables. This heatmap is very useful in the selection of variables for building Machine Learning models.

Variables that have a good relationship with ‘Price’ can be used for predicting price.

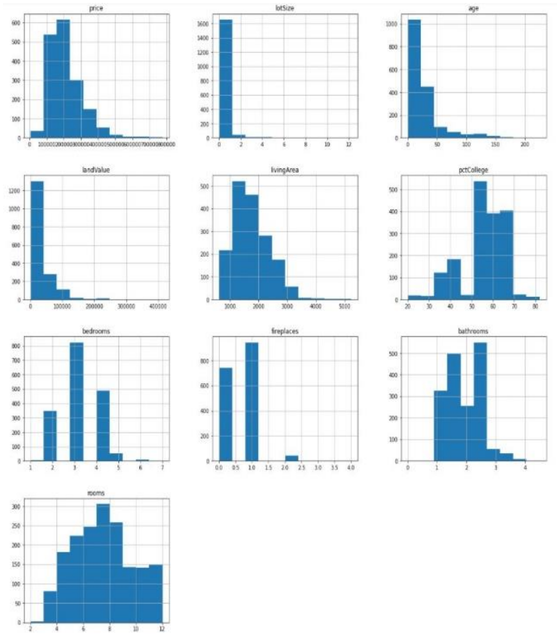
From the below heatmap, we can conclude that Living Area, Bathrooms, Land value and rooms are highly correlated with Price compared to the other independent variables.



Histogram shows the probability distribution. Histogram graphically shows the accurate distribution of numerical data.

Following information’s are extracted from each variable’s graph in Histogram:

- ‘price’ is normally distributed. There are a few outliers that have very low and very high price.
- ‘lotSize’ shows that majority of houses have a lot size less than 2.
- ‘age’ is not normally distributed. This graph shows that there are few houses which are very old i.e., 100+ years. But from the ‘Age’ graph, we can conclude that most of the houses are new and have less than 50 years of age.
- ‘landValue’ shows the value of land where the house was built. This shows that most of the houses are inland whose value is less than 100,000.
- ‘livingArea’ shows that space of living area in houses i.e. the majority of houses have a living area less than 3500.
- ‘pctCollege’ shows that 50-70% as densely populated. That means around 50-70% of neighbours have graduated from college.
- ‘bedrooms’ shows the numbers of bedrooms. This graph shows that majority of Houses have 3 bedrooms than 4 bedrooms and 2 bedrooms. There are very few houses with 1, 5 or 6 bedrooms.
- ‘firplaces’ shows that most of the houses have 0 or 1 fireplace but there are very few houses that have more than 2 fireplaces.
- ‘bathrooms’ shows that majority of houses have 1,2 or 3 bathrooms and there are very few houses with 3 or more bathrooms.
- ‘rooms’ shows that majority of houses have 6 to 9 rooms in total. However, there are few houses with more than 9 rooms and also less than 6 rooms.



IV. MODEL BUILDING STEPS:

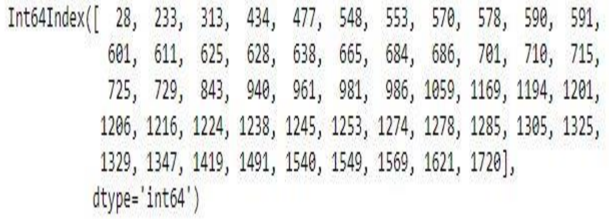
Treatment of outliers is very important to get much more accuracy.

Upper and lower outliers can be found using below mentioned formula:

Upper -> Q3 + 1.5 * IQR

Lower -> Q1 – 1.5 * IQR

Below are the outliers:



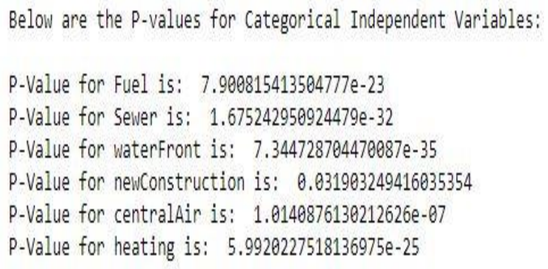
```
Int64Index([  28,  233,  313,  434,  477,  548,  553,  570,  578,  590,  591,
             601,  611,  625,  628,  638,  665,  684,  686,  701,  710,  715,
             725,  729,  843,  940,  961,  981,  986, 1059, 1169, 1194, 1201,
            1206, 1216, 1224, 1238, 1245, 1253, 1274, 1278, 1285, 1305, 1325,
            1329, 1347, 1419, 1491, 1540, 1549, 1569, 1621, 1720],
           dtype='int64')
```

Outliers will be stored in a new ‘DataFrame’.

Now, dropped the Outliers from the main ‘DataFrame’, that we are going to use for model building.

Numeric values of dataset will be scaled using sklear to transform the values and create a new ‘DataFrame’ using the transformed values of Independent variables.

As there are few Categorical Variables in the ‘DataFrame’, so we need to check whether those Categorical Variables will be useful in predicting price or not, for that, I used ‘One Way Anova Test’ to get P-values. Now, if P-Value > 0.05, we accept the H0 and if P-Value < 0.05 then we can fail to accept HO. We will only use Categorical Variables if P-Value < 0.05.



```
Below are the P-values for Categorical Independent Variables:

P-Value for Fuel is:  7.900815413504777e-23
P-Value for Sewer is:  1.675242950924479e-32
P-Value for waterFront is:  7.344728704470087e-35
P-Value for newConstruction is:  0.031903249416035354
P-Value for centralAir is:  1.0140876130212626e-07
P-Value for heating is:  5.9920227518136975e-25
```

Now, if we want to use Categorical Variables in Linear Regression model, then we need to convert those Categorical Variables into ‘Dummies’.

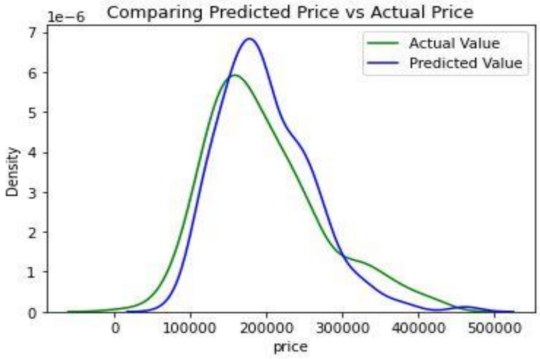
Dummy variables are used and dropped actual Categorical Variables from the ‘DataFrame’, so that we only have numerical values left, which will help in predicting price.

train_test_split is used from sklear.model_selection and divided the data into train and test. Where, 80% of data is for training and 20% of data is for testing the model.

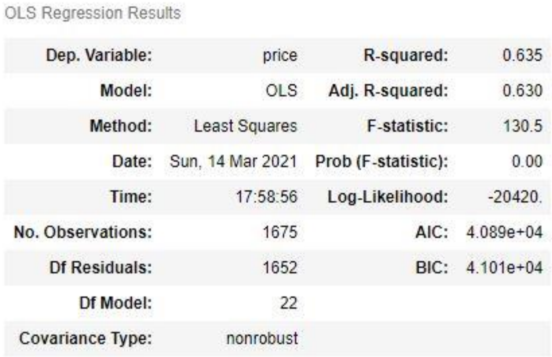
Used x_train and y_train to fit LinearRegression and from this we get our prediction of price.

After getting the predicted price list, sklearn.metrics is used to find R-sqr value. After comparing Prediction with Actual price, observed R-sqr was 0.6084320849535112, which means our model has 60.84% accuracy in prediction of price.

Below is the distplot graph which demonstrates comparison of Predicted Price vs Actual Price.



Moreover, to improve our prediction, we need to use ‘StatsModel’ to compute the ‘OLS Regression Results’. Here, the Adj, R-sqr is 0.630, i.e., 63% prediction.



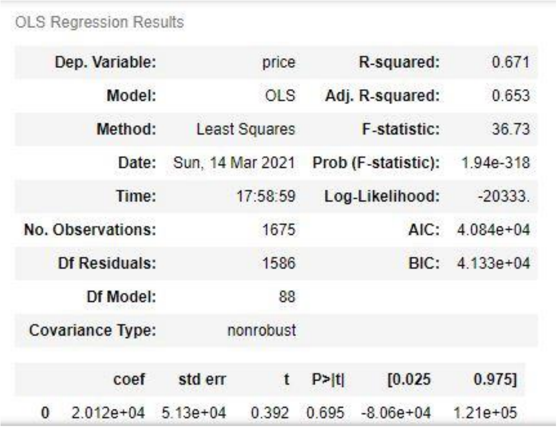
OLS Regression Results

Dep. Variable:	price	R-squared:	0.635
Model:	OLS	Adj. R-squared:	0.630
Method:	Least Squares	F-statistic:	130.5
Date:	Sun, 14 Mar 2021	Prob (F-statistic):	0.00
Time:	17:58:56	Log-Likelihood:	-20420.
No. Observations:	1675	AIC:	4.089e+04
Df Residuals:	1652	BIC:	4.101e+04
Df Model:	22		
Covariance Type:	nonrobust		

We will also get list of all Independent Variables in OLS Regression Results. We can check coefficient, standard error, t value and p value. From this table we can check whether to use Independent variable or not. If P > 0.05 then that variable is not useful and if P < 0.05 then that means, variable is helpful in prediction.

Now, Forward and Backward feature selector are used to get the best variables that we need. A dictionary of those variables which was used is created and now a new DataFrame of those variables is created.

Again, ‘StatsModel’ is used to get the ‘OLS Regression Results’.



OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.671
Model:	OLS	**Adj. R-squared:**	0.653
Method:	Least Squares	**F-statistic:**	36.73
Date:	Sun, 14 Mar 2021	**Prob (F-statistic):**	1.94e-318
Time:	17:58:59	**Log-Likelihood:**	-20333.
No. Observations:	1675	**AIC:**	4.084e+04
Df Residuals:	1586	**BIC:**	4.133e+04
Df Model:	88		
Covariance Type:	nonrobust		

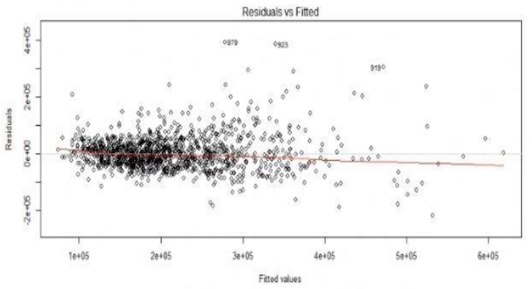
	coef	std err	t	P>|t|	[0.025	0.975]
0	2.012e+04	5.13e+04	0.392	0.695	-8.06e+04	1.21e+05

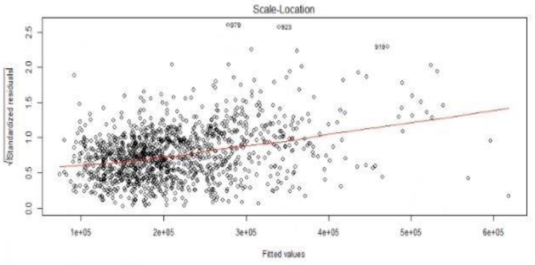
From the above ‘OLS Regression Results’, we can find the Adj R-sqr 0.653, i.e., 65.3% prediction accuracy.

V. GAUSS MARKOV ASSUMPTION:

According to Gauss Markov, there will be no systematic relation between the Residuals and Predicted values if both dependent and independent variables are linearly related.

Model will leave random noise and will take all the systematic variance present in the data.





Linearity: In the above plot of Residual vs Fitted does not have any systematic pattern. All points are randomly scattered. Hence, Linearity assumption is satisfied.

Homscedasticity: No systematic pattern is observed in the plot of Residuals and Fitted values. All points are randomly scattered. Hence, Homscedasticity assumption is satisfied.

VI. SUMMARY OF THE FINAL MODEL:

In the starting we used ‘Linear Regression’. The accuracy was 60%. After that ‘StatsModel’ is used and OLS Results has 63% accuracy.

Next, we tried to find best ‘Independent Variables’ using ‘Forward and Backward selection’. Then again, using those ‘Independent Variables’, when those variables are used for ‘StatsModel’, we got the OLS Results that has 65% accuracy.

Conclusively, we can say when we compared our accuracy of different models. The best that we got is 65%. We used ‘StatsModel’ for that and used best Independent Variables. This OLS Regression Results also shows the coefficient of each Individual Variables and by checking P-values, we can understand the impact of independent variables on our model.

Final model is not overfitting training data and model is generalized.

VII. REFERENCES:

[1] Statistics and Machine Learning in

Python by Edouard Duchesnay, Tommy Löfstedt

[2] Understanding and Using Advanced Statistics SAGE, p.178, [ISBN: 141290014X]

[3] Predicting Sales Prices of the Houses Using

Regression Methods of Machine Learning - IEEE

Paper.