Housing pricing prediction model

Overview

Since real estate investing is one of the top performing industries globally, there is a lot of interest in figuring out what factors should be taken into account before making a real estate investment. As a result, this study goes further to investigate the variables that drive housing prices to fluctuate, a choice that has an impact on stakeholders and homebuyers or tenants. Because of this, house price modelling will assist both stakeholders and tenants in making wise judgments and in determining the precise costs associated with investing in real estate or the price of a home, regardless of the services offered. As a result, this will benefit everyone, including stakeholders, as well as the government, which will benefit from knowing how the economy is doing.

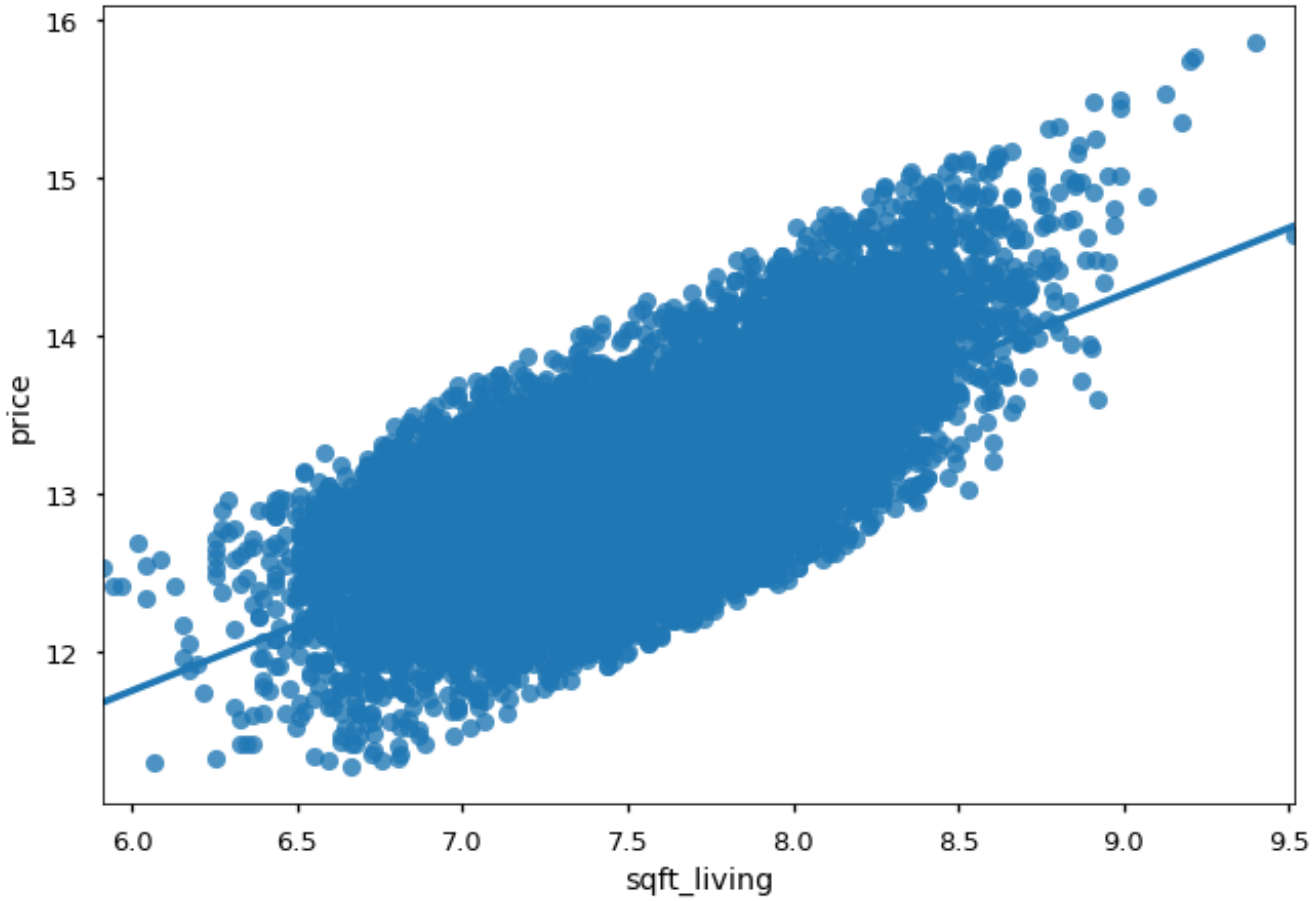
Business Understanding

Here, a data scientist's responsibility is to find out what factors are most likely to have an impact on home prices. They also look closely at how the variables affect home prices and how much they will impact the stakeholders in the housing market. This is really important because before real estate investors make a decision, they need think about whether the investment will be profitable or not. Data scientists are consequently required to conduct important evaluations and analyses to identify the crucial variables and their influence on investment decisions. Instead of repeatedly going through the process, after careful study, a predictive model that will be utilized to assist new investors entering the market should be constructed.

Data Understanding and Analysis

Secondary data from King County House Sales has been employed to be used in this project as "kc_house_data.scv".

The most correlated factor with house pricing



Simple Linear Regression

OLS Regression Results

Dep. Variable:	price	R-squared:	0.455
Model:	OLS	Adj. R-squared:	0.455
Method:	Least Squares	F-statistic:	1.805e+04
Date:	Fri, 30 Sep 2022	Prob (F-statistic):	0.00
Time:	13:44:48	Log-Likelihood:	-10231.
No. Observations:	21597	AIC:	2.047e+04
Df Residuals:	21595	BIC:	2.048e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.7234	0.047	142.612	0.000	6.631	6.816
sqft_living	0.8376	0.006	134.368	0.000	0.825	0.850

Omnibus:	123.577	Durbin-Watson:	1.977
Prob(Omnibus):	0.000	Jarque-Bera (JB):	114.096
Skew:	0.143	Prob(JB):	1.68e-25
Kurtosis:	2.787	Cond. No.	137.

Interpretation of Simple Linear Regression

sqft_living was the attribute most strongly correlated with price, therefore our model is describing this relationship.

Overally, this model is statistically significant because Prob (F-statistic):0.00 is less than 0.05(alpha) and R squared is 45.5% which explains about 45.5% of the variance in price indicating the other percentage is explained by the error.

Also,the coefficient is statistically significant because the p value = 0.00 is less than our level of significance

In a typical explanation,it means that a unit increase in sqft_living by one unit, will make the price of the house to increase by by 0.8376%,this is because coefficients are interpreted using percentage when they are log transformed.

The intercept is at about 1.4418% , without considering sqft_living15 the price of houses is likely to be 1.4418% .

Multiple Linear Regression

Dep. Variable:	price	R-squared:	0.506
Model:	OLS	Adj. R-squared:	0.506
Method:	Least Squares	F-statistic:	2455.
Date:	Fri, 30 Sep 2022	Prob (F-statistic):	0.00
Time:	19:45:09	Log-Likelihood:	-9181.8
No. Observations:	21597	AIC:	1.838e+04
Df Residuals:	21587	BIC:	1.846e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.5706	0.059	93.935	0.000	5.454	5.687
sqft_living	0.6317	0.013	48.783	0.000	0.606	0.657
sqft_living15	0.4729	0.012	38.877	0.000	0.449	0.497
sqft_above	-0.0070	0.012	-0.567	0.571	-0.031	0.017
sqft_lot	-0.0195	0.007	-2.728	0.006	-0.034	-0.006
sqft_lot15	-0.0736	0.008	-9.267	0.000	-0.089	-0.058
condition_Fair	-0.1173	0.029	-4.076	0.000	-0.174	-0.061
condition_Good	0.0611	0.006	10.118	0.000	0.049	0.073
condition_Poor	-0.1361	0.069	-1.972	0.049	-0.271	-0.001
condition_Very Good	0.1869	0.010	19.279	0.000	0.168	0.206

Interpretation of Multiple Linear Regression

The model is statistically significant with all the coefficients, this is because they have a p value less than the level of siginificance except sqft_above and sqft_lot. The model has R_squared of 50.6%, indicating that our dependent variables is explained 50.6% of the variance explained by the model.

sqft_living and sqft_living15 affect the model positively, that is when there is a one unit percentage increase holding all the other explanatory variables constant where sqft_living will make price to change by 0.6317% holding all other explanatory variable constant and sqft_living15 will make the price to change by 0.4729% and all the other explanatory variables makes price to decrease because they have a negative coeffients.

Interpretation of condition_fair, condition_Good, condition_poor and condition_very Good they are not same with the one above because their effect are comparable to condition_Average. Typically condition condition_poor and condition_fair are likely to have a negative effect on the prediction of prices because condition_Average is higher than the two,thus reason for having negative coefficients,condition_Good and condition_Very Good have a positive impact on condition average making it to influence price positively.

The effect change of condition_Very Good is 0.1869 on condition_Average indicating it has more effect and the effect change of condition_Good on condition_Average is 0.0611 indicating it weighs more higher effect on price. The effect of condition_poor is -0.1361 on condition_Average thus it indicates it has a reducing effect on pricing of houses and the effect of condition_Fair is -0.1173 indicating it has a negative impact on condition_Average.

Final predictive model:

price = 5.5706 + 0.6317sqft_living+ 0.4729sqft_living15 - 0.0736sqft_lot15 - 0.1173condition_Fair + 0.0611condition_Good - 0.1361condition_Poor + 0.1869condition_Very Good