King County Housing Case Study

(Factors that Influence Property Sale Prices)

Ravinarayan Raghupathi, MSc PhD

Overview

The brief

Use multiple linear regression modeling to analyse house sales in a northwestern US county.

Business problem

Which factors influence and can help improve King County house sales?

Strategy

Examine the following key indicators (independent variables) from the given dataset that affect the sale prices of properties (the dependent variable) in King County:

- Number of bedrooms
- Number of bathrooms
- Living area (in square feet)
- Lot area (in square feet)
- Number of floors
- Condition of the property
- Property grade

Create another independent variable which is the age of the property

Methodology

Exploratory Data Analysis

⬇

Model Iteration 1 (test assumptions of regression)

⬇

Transform independent and categorical variables

⬇

Model Iteration 2 (test assumptions of regression)

⬇

Any further transformation of data and new Model iteration(s), and test assumptions of regression

⬇

Final model validation and conclusions

Model 1 (Baseline)

The baseline model was created using all the available independent variables as is, without any transformations or manipulation.

Observations

- The model appeared to indicate some linear relationship between the chosen independent and dependent variables.
- There was a combination of continuous and categorical variables.
- With the exception of 'Age', the distribution of all continuous variables appeared to be heavily right-skewed, probably because of outliers, which will need to be eliminated.
- The assumptions of linear regression could not be satisfied in this iteration.

Model 2 (Iteration 2)

This model was created after eliminating outliers, dealing with categorical variables, transforming the independent variables and checking for multicollinearity.

Observations

- Despite eliminating variables that were shown to be multicollinear, there was still a suggestion of multicollinearity (high Condition Number).
- The skew and kurtosis values were closer to values representing normal distribution.
- There were a few categorical variables whose p-values indicated that they were not significant and could be eliminated in the next iteration.
- The assumptions of linear regression had improved and suggested that further tweaking of the variables might lead to a better outcome.

Model 4 (Iteration 4)

The final model (Model 4) was created after eliminating non-significant variables in Model 3 (Iteration 3) and then performing feature scaling on the independent variables.

Observations

- Multicollinearity and other errors were eliminated by removing non-significant variables.
- There was no significant change in the assumptions of regression from Model 2 but the relationship between the independent and dependent variables were still reasonably robust.
- The coefficients for 'Living area' and 'Age' were positive, indicating that an increase in either caused an increase in sale price, whilst the opposite was true of 'Lot area' (with a negative coefficient).

Comparison of Models

OLS Regression Results

Dep. Variable:	Price	**R-squared:**	0.618
Model:	OLS	**Adj. R-squared:**	0.618
Method:	Least Squares	**F-statistic:**	4363.
		Prob (F-statistic):	0.00
		Log-Likelihood:	-2.9700e+05
No. Observations:	21597	**AIC:**	5.940e+05
Df Residuals:	21588	**BIC:**	5.941e+05
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.103e+06	1.82e+04	-60.519	0.000	-1.14e+06	-1.07e+06
Bedrooms	-4.915e+04	2123.053	-23.151	0.000	-5.33e+04	-4.5e+04
Bathrooms	5.286e+04	3587.694	14.734	0.000	4.58e+04	5.99e+04
Living_area	187.4021	3.421	54.784	0.000	180.697	194.107
Lot_area	-0.2459	0.038	-6.439	0.000	-0.321	-0.171
Floors	2.128e+04	3592.816	5.922	0.000	1.42e+04	2.83e+04
Condition	1.962e+04	2583.883	7.593	0.000	1.46e+04	2.47e+04
Grade	1.311e+05	2238.758	58.577	0.000	1.27e+05	1.36e+05
Age	4010.7386	69.171	57.983	0.000	3875.159	4146.318

Omnibus:	17302.265	**Durbin-Watson:**	1.984
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1207162.645
Skew:	3.353	**Prob(JB):**	0.00
Kurtosis:	39.007	**Cond. No.**	5.24e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.24e+05. This might indicate that there are strong multicollinearity or other numerical problems.

OLS Regression Results

Dep. Variable:	Price	**R-squared:**	0.539
Model:	OLS	**Adj. R-squared:**	0.538
Method:	Least Squares	**F-statistic:**	793.5
		Prob (F-statistic):	0.00
		Log-Likelihood:	-2.3439e+05
No. Observations:	17703	**AIC:**	4.688e+05
Df Residuals:	17676	**BIC:**	4.690e+05
Df Model:	26		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.272e+06	5.64e+04	-22.565	0.000	-1.38e+06	-1.16e+06
Living_area	2.428e+05	5363.035	45.279	0.000	2.32e+05	2.53e+05
Lot_area	-6.957e+04	2331.867	-29.836	0.000	-7.41e+04	-6.5e+04
Age	1.078e+05	2398.003	44.973	0.000	1.03e+05	1.13e+05
Bed_3	-4.237e+04	3519.801	-12.036	0.000	-4.93e+04	-3.55e+04
Bed_4	-5.042e+04	4272.505	-11.800	0.000	-5.88e+04	-4.2e+04
Bed_5	-4.962e+04	5924.508	-8.375	0.000	-6.12e+04	-3.8e+04
Bath_1.0	336.5176	3898.211	0.086	0.931	-7304.358	7977.393
Bath_1.5	-1.828e+04	4339.078	-4.212	0.000	-2.68e+04	-9770.599
Bath_2.0	-8286.3994	3820.978	-2.169	0.030	-1.58e+04	-796.908
Bath_2.5	-1.506e+04	3079.291	-4.892	0.000	-2.11e+04	-9027.381
Bath_3.0	2546.6310	6352.150	0.401	0.688	-9904.207	1.5e+04
Bath_3.5	5.455e+04	7321.492	7.450	0.000	4.02e+04	6.89e+04
Flr_1.5	2.16e+04	4045.951	5.339	0.000	1.37e+04	2.95e+04
Flr_2.0	2539.4420	3370.846	0.753	0.451	-4067.747	9146.631
Flr_2.5	2.277e+04	1.51e+04	1.509	0.131	-6801.002	5.23e+04
Flr_3.0	3.41e+04	7486.617	4.554	0.000	1.94e+04	4.88e+04
Flr_3.5	1.989e+04	5.59e+04	0.356	0.722	-8.96e+04	1.29e+05
Cond_2	-4.788e+04	1.26e+04	-3.814	0.000	-7.25e+04	-2.33e+04
Cond_5	4.297e+04	3948.579	10.883	0.000	3.52e+04	5.07e+04
Grd_5	-9308.5038	4.44e+04	-0.210	0.834	-9.64e+04	7.78e+04
Grd_6	2.691e+04	4.33e+04	0.621	0.534	-5.8e+04	1.12e+05
Grd_7	1.006e+05	4.33e+04	2.324	0.020	1.57e+04	1.85e+05
Grd_8	2.035e+05	4.34e+04	4.689	0.000	1.18e+05	2.89e+05
Grd_9	3.548e+05	4.36e+04	8.136	0.000	2.69e+05	4.4e+05
Grd_10	4.415e+05	4.4e+04	10.026	0.000	3.55e+05	5.28e+05
Grd_11	5.333e+05	4.76e+04	11.213	0.000	4.4e+05	6.26e+05

Omnibus:	879.812	**Durbin-Watson:**	1.983
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1262.180
Skew:	0.461	**Prob(JB):**	8.34e-275
Kurtosis:	3.928	**Cond. No.**	1.46e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.46e+03. This might indicate that there are strong multicollinearity or other numerical problems.

OLS Regression Results

Dep. Variable:	Price	**R-squared:**	0.537
Model:	OLS	**Adj. R-squared:**	0.536
Method:	Least Squares	**F-statistic:**	1206.
		Prob (F-statistic):	0.00
		Log-Likelihood:	-2.3442e+05
No. Observations:	17703	**AIC:**	4.689e+05
Df Residuals:	17685	**BIC:**	4.690e+05
Df Model:	17		
Covariance Type:	nonrobust		

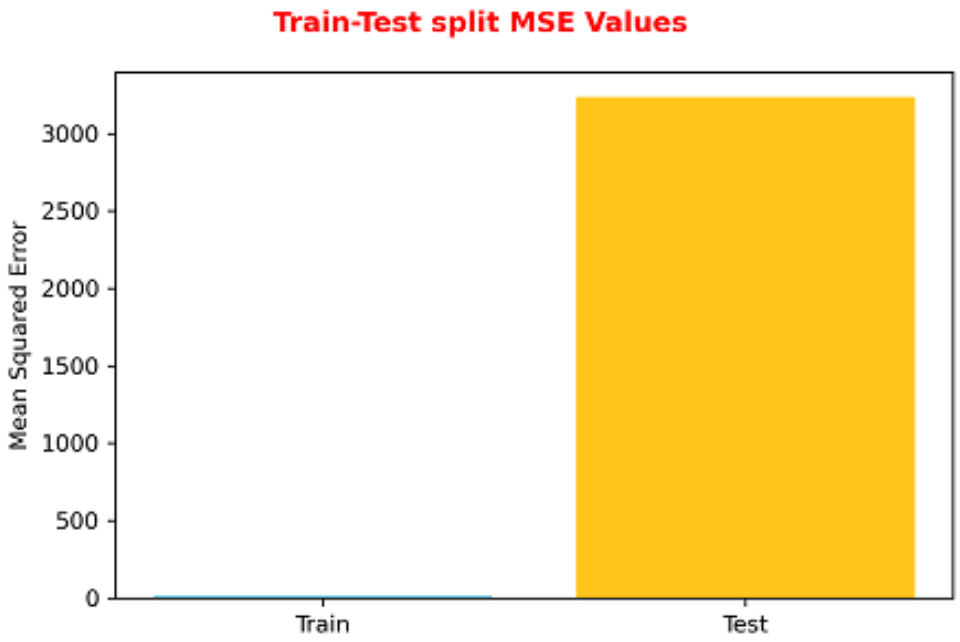
	coef	std err	t	P>|t|	[0.025	0.975]
const	3.671e+05	4811.177	76.307	0.000	3.58e+05	3.77e+05
Bed_3	-3.898e+04	3456.235	-11.279	0.000	-4.58e+04	-3.22e+04
Bed_4	-4.662e+04	4210.764	-11.073	0.000	-5.49e+04	-3.84e+04
Bed_5	-4.551e+04	5867.379	-7.757	0.000	-5.7e+04	-3.4e+04
Bath_1.5	-1.758e+04	4045.778	-4.346	0.000	-2.55e+04	-9651.724
Bath_2.0	-8147.7104	3600.259	-2.263	0.024	-1.52e+04	-1090.849
Bath_2.5	-1.591e+04	2888.735	-5.507	0.000	-2.16e+04	-1.02e+04
Bath_3.5	5.32e+04	7240.209	7.347	0.000	3.9e+04	6.74e+04
Cond_2	-4.971e+04	1.26e+04	-3.957	0.000	-7.43e+04	-2.51e+04
Cond_5	4.313e+04	3938.708	10.951	0.000	3.54e+04	5.09e+04
Grd_7	7.563e+04	3811.218	19.844	0.000	6.82e+04	8.31e+04
Grd_8	1.797e+05	4595.740	39.093	0.000	1.71e+05	1.89e+05
Grd_9	3.312e+05	5900.052	56.136	0.000	3.2e+05	3.43e+05
Grd_10	4.189e+05	8353.619	50.140	0.000	4.02e+05	4.35e+05
Grd_11	5.12e+05	1.98e+04	25.905	0.000	4.73e+05	5.51e+05
Living_area	5.547e+05	1.1e+04	50.442	0.000	5.33e+05	5.76e+05
Lot_area	-2.762e+05	7228.360	-38.212	0.000	-2.9e+05	-2.62e+05
Age	3.001e+05	5627.064	53.327	0.000	2.89e+05	3.11e+05

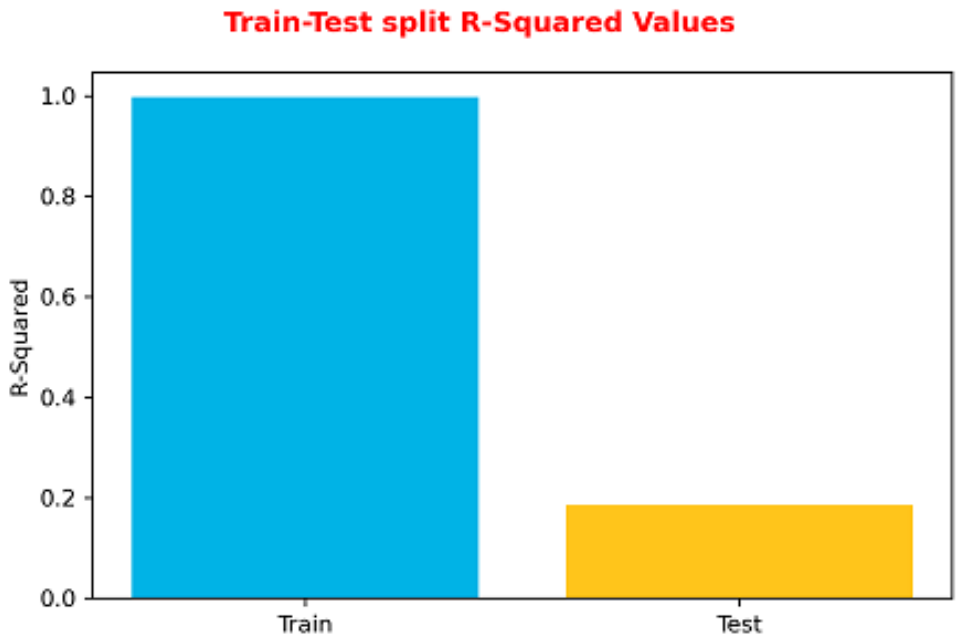
Omnibus:	877.248	**Durbin-Watson:**	1.982
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1255.056
Skew:	0.461	**Prob(JB):**	2.94e-273
Kurtosis:	3.923	**Cond. No.**	26.5

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Model validation

The final model (Model 4) was evaluated using Train-Test splits and the prediction accuracy was computed.





Computed accuracy of the model: 0.5400669334410992

Conclusions

1. The final model is overfitted and will predict correctly around 54% of the time, which is acceptable since it is an inference model.
2. The best indicator for a good sale price for a property appears to be the size of the living area. The age of the property might be a factor.
3. The area of the lot, and the number of bedrooms and bathrooms has an inverse relationship with sale prices.

Actionable insight

The two main factors I would choose to infer property sale prices in King County are ‘Living area' and ‘Lot area'.