King County

House Pricing Analysis and Predictive Modelling

By Mitra Zamani

July 2, 2023

Summary

- For this project, I will use regression modelling to analyse house sales in a North Western county. This project will also build pricing models that will help to best price houses based on their features.

Outline:

- **Business Problem**
- **Data and Methods**
- **Results**
- **Conclusions**

Business Problem:

- I am tasked to help a Real Estate Buyer's Agency in seeking to identify key property features, such as square footage and number of rooms, to identify undervalued properties that can be presented as investment opportunities for clients.

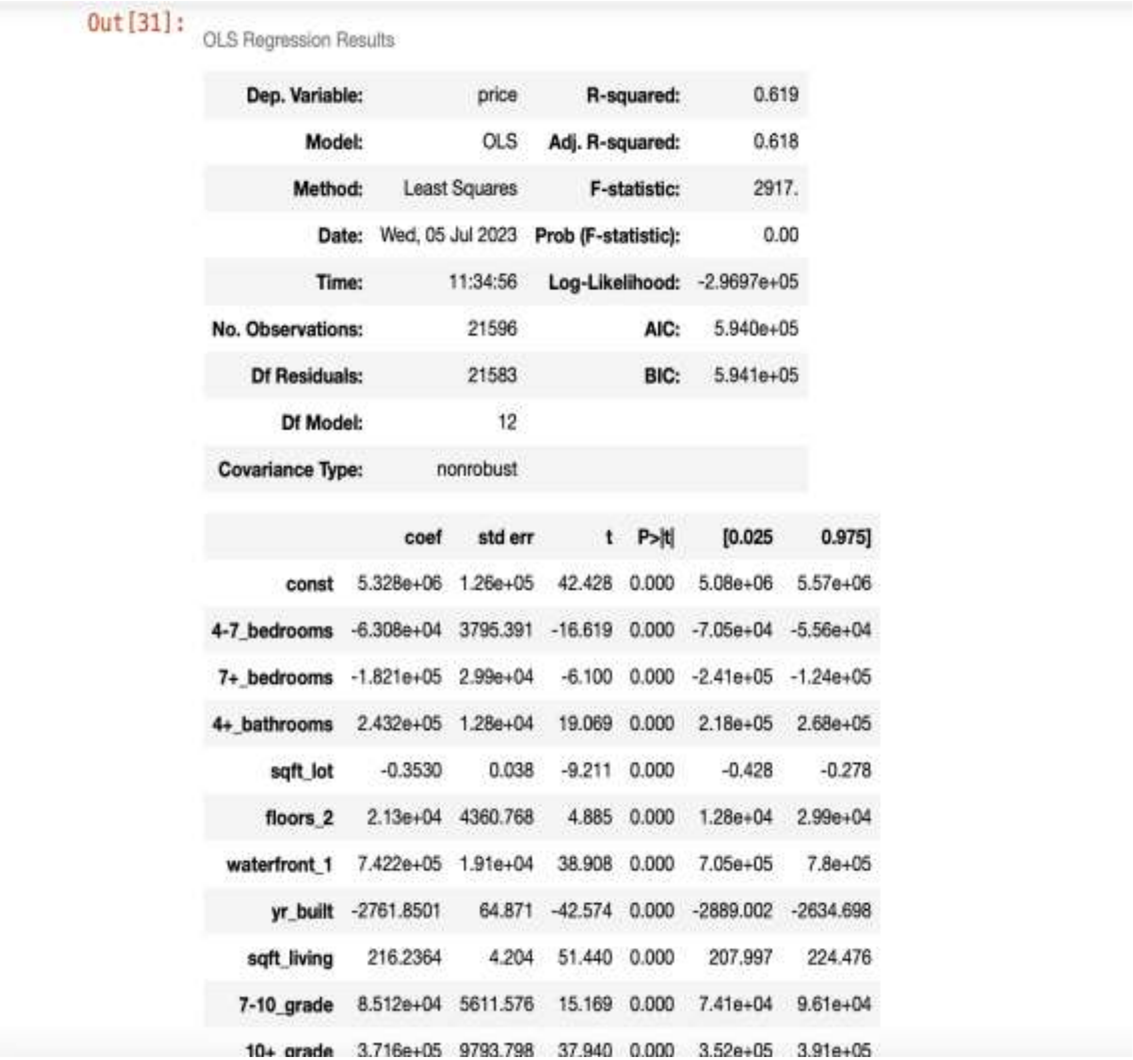
Data and Methods:

- The data utilized in this project was sourced from the kc_house_data CSV file. The data set is comprised of 21,597 rows and 21 columns, providing an ample amount of information for modelling purposes. This comprehensive data set is well-suited to support the modelling process and can be expected to yield reliable results.

- This project will employ the 'OSEMN' data science process to source the data, perform exploratory data analysis, clean and prepare the data for model, fit regression models, Interpret results that will be used to make recommendations to the firm.

Results:

- **Iteration 1:**
- Baseline Model:
- After cleaning Data and some changes on categorical data(create dummies), I created Baseline model. The model’s performance was not very bad with an r-squared of 0.619 , explaining 62% of the variance.



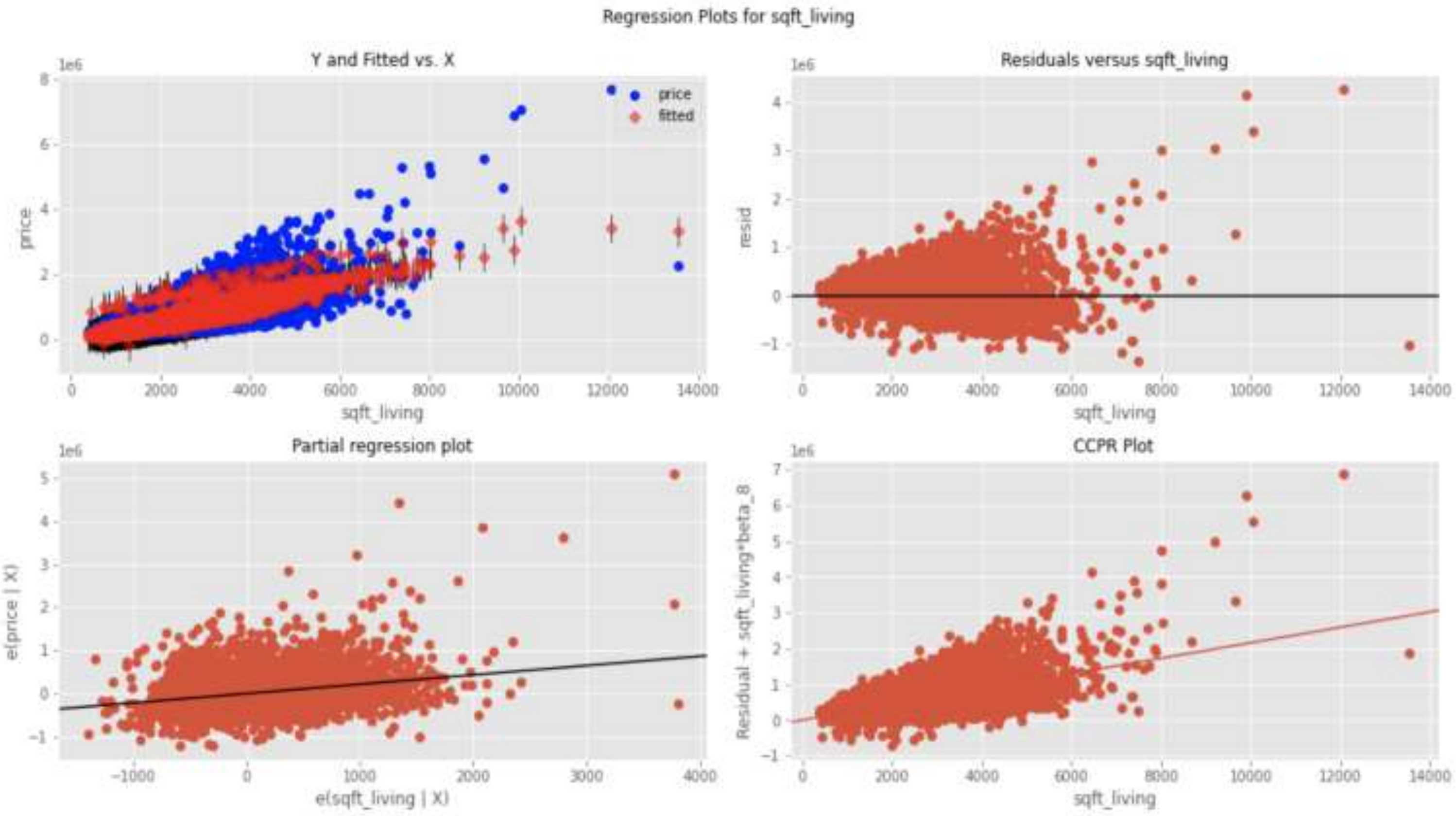
Out[31]:

OLS Regression Results

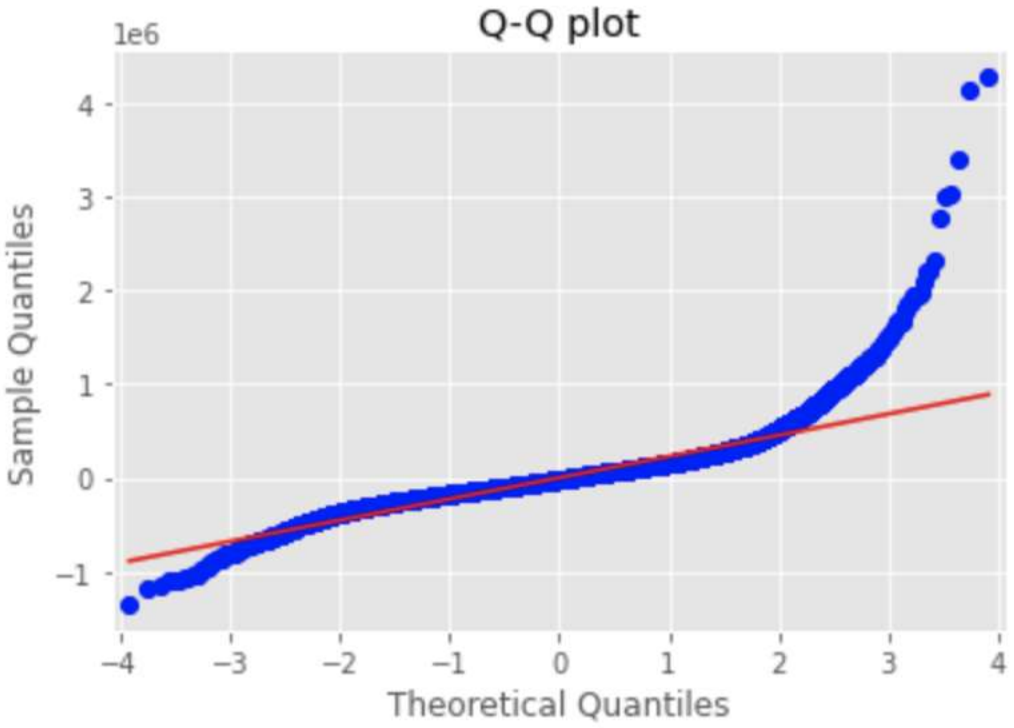
Dep. Variable:	price	**R-squared:**	0.619
Model:	OLS	**Adj. R-squared:**	0.618
Method:	Least Squares	**F-statistic:**	2917.
Date:	Wed, 05 Jul 2023	**Prob (F-statistic):**	0.00
Time:	11:34:56	**Log-Likelihood:**	-2.9697e+05
No. Observations:	21596	**AIC:**	5.940e+05
Df Residuals:	21583	**BIC:**	5.941e+05
Df Model:	12		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	5.328e+06	1.26e+05	42.428	0.000	5.08e+06	5.57e+06
4-7_bedrooms	-6.308e+04	3795.391	-16.619	0.000	-7.05e+04	-5.56e+04
7+_bedrooms	-1.821e+05	2.99e+04	-6.100	0.000	-2.41e+05	-1.24e+05
4+_bathrooms	2.432e+05	1.28e+04	19.069	0.000	2.18e+05	2.68e+05
sqft_lot	-0.3530	0.038	-9.211	0.000	-0.428	-0.278
floors_2	2.13e+04	4360.768	4.885	0.000	1.28e+04	2.99e+04
waterfront_1	7.422e+05	1.91e+04	38.908	0.000	7.05e+05	7.8e+05
yr_built	-2761.8501	64.871	-42.574	0.000	-2889.002	-2634.698
sqft_living	216.2364	4.204	51.440	0.000	207.997	224.476
7-10_grade	8.512e+04	5611.576	15.169	0.000	7.41e+04	9.61e+04
10+_grade	3.716e+05	9793.798	37.940	0.000	3.52e+05	3.91e+05

- After creating model I Verified the Assumptions of Linear Regression. From the first and second plot in the first row, we see a cone-shape which is a sign of heteroscedasticity. i.e. the residuals are heteroscedastic. This violates an assumption.



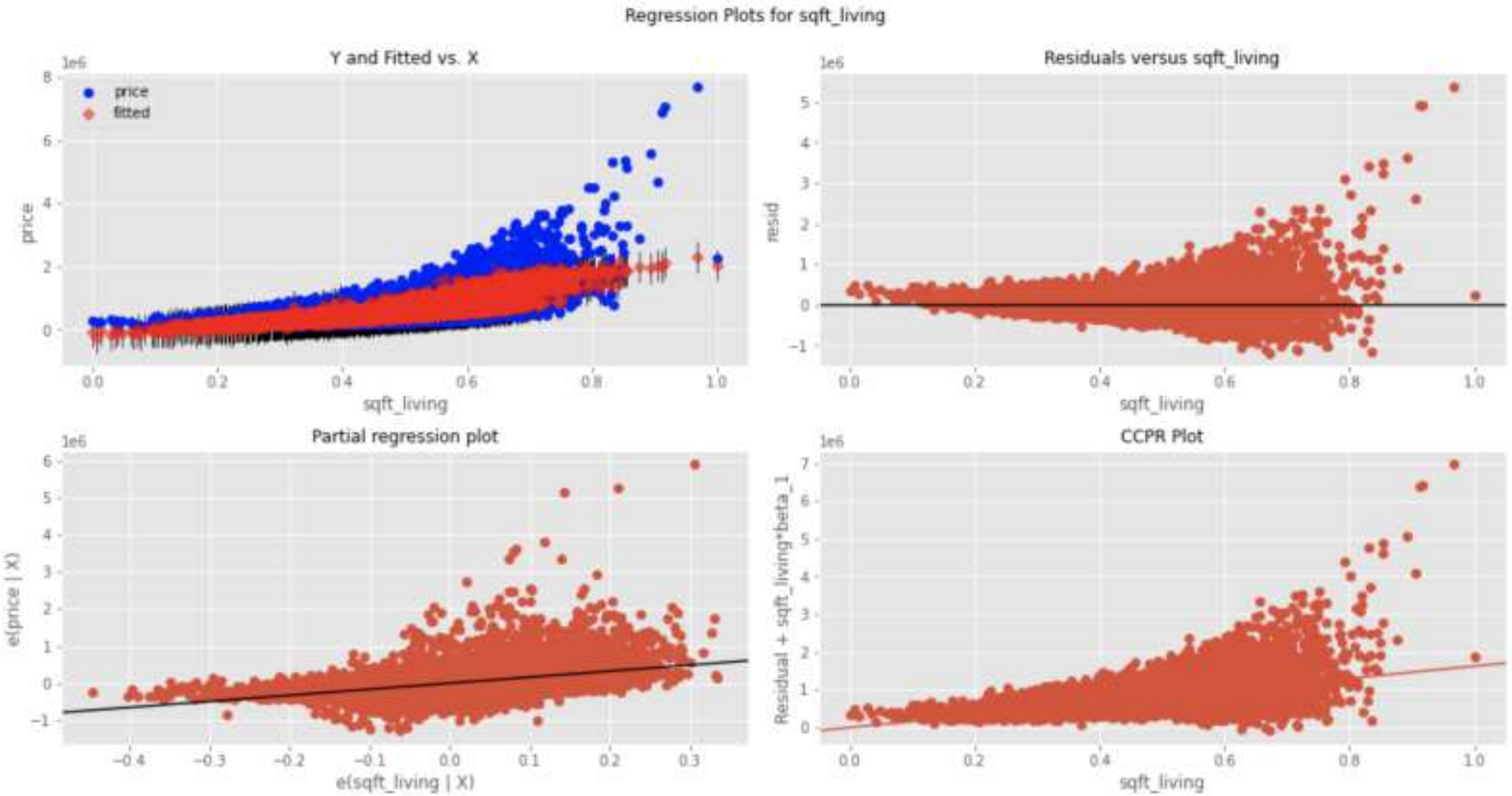
- Then I plotted Q-Q plot to check the normality of residuals.
- Notice the points fall along a line in the middle of the graph, but curve off in the extremities. Normal QQ plots that exhibit this behaviour usually mean your data have more extreme values than would be expected if they truly came from a normal distribution.



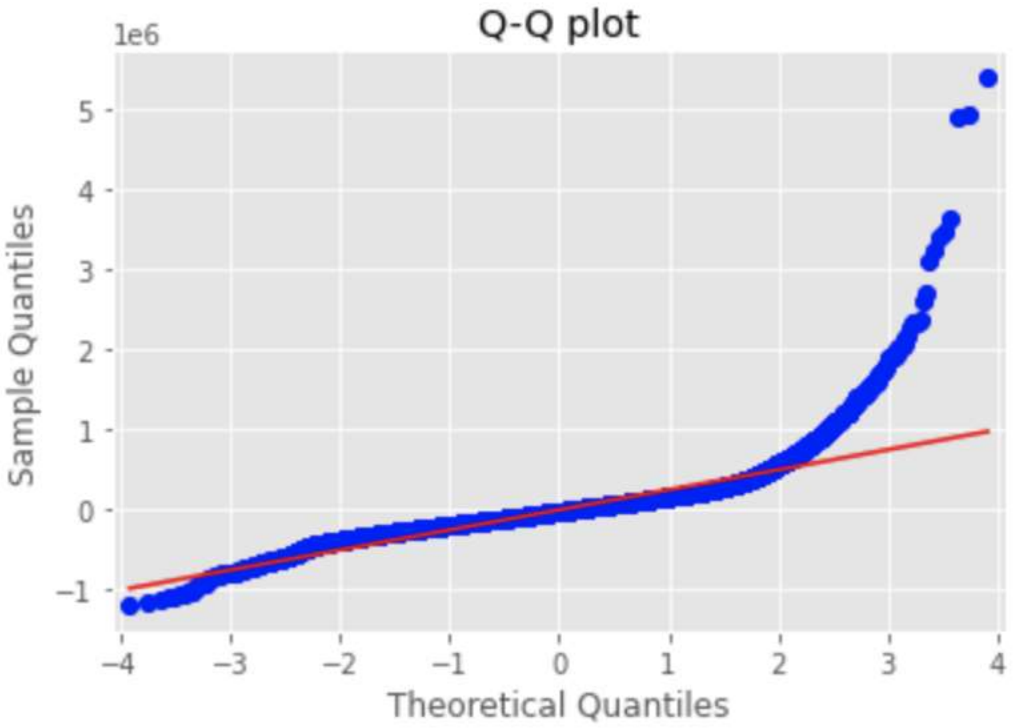
Iteration 2

- Model 2:
- I identified multicollinearity by correlation matrix plot, and removed the desired variables.
- Also I removed any other variables which didn’t make sense to go in the model.
- I checked the normality of continues data and then I did Log Transformations for reduce the skewness.
- After that I created my second model. The model’s performance was not very good with an r-squared of 0.535 , explaining 54% of the variance.

- After creating model I Verified the Assumptions of Linear Regression. From the first and second plot in the first row, figures illustrate a heteroscedastic data set. i.e. the residuals are heteroscedastic. This violates an assumption.



- Then I plotted Q-Q plot to check the normality of residuals.
- Notice the points fall along a line in the middle of the graph, but curve off in the extremities. Normal QQ plots that exhibit this behaviour usually mean your data have more extreme values than would be expected if they truly came from a normal distribution. the residuals QQ plot looks off, so the normality assumption is not fulfilled.



Iteration 3

- I did more transformations. And then I created my model again.
- The model's performance is same as the second model with an r-squared of 0.535 , explaining 54% of the variance.
- I Verified the Assumptions of Linear Regression and plotted Q-Q plot. Both of them were same as the second model.
- Then I did Model validations (Train/Test Split) and Calculated the Mean Squared Error (MSE). Train MSE is higher than Test MSE, so our model is underfitting. When a model has not learned the patterns in the training data well and is unable to generalize well on the new data, it is known as underfitting. An underfit model has poor performance on the training data and will result in unreliable predictions.

Conclusion:

- **Findings and Recommendations:**
- The overall living size area of the house i.e. excluding the basement is very significant to the price. The study show a linear relationship between the size of the living space and price while the size of the basement played a very insignificant role to the houses' selling price. My recommendation would therefore be to acquire houses with a relatively larger living space as compared to the basement.

Future Work:

- Model three couldn’t be our final model because the r-squared not improved from iteration 2 and also after I did model validation, I realized the model is underfitting. So for the next step I should do one of these to tackle underfitting.
- Increase the number of features in the dataset
- Increase model complexity
- Reduce noise in the data
- Increase the duration of training the data

In this case, the next steps would be reducing noise in the data to improve the accuracy of our model. Additionally, i would like to investigate certain features like proximity to good schools, other facilities like hospital, gyms, restaurants and play grounds.

Thank you!

Email: zmitra15@gmail.com
GitHub: @MitraZamani