# Housing in King County

Anh Luong

# Problem

This presentation was prepared after we analysed the housing data in King County.

This purpose of this analysis is to determine which features of a property are critical to a property pricing.

# What did we look at?

- **Sales data of King County**
- **Multiple features of a property**
  - Number of Bedrooms
  - Number of bathrooms
  - Footage of the home
  - Footage of the lot
  - A view to a waterfront
  - Condition - How good the condition is ( Overall )
  - Grade - overall grade given to the housing unit, based on King County grading system
  - Age

# What are the most important features?

# Summary

R-square - 0.516

4 significant variables based on P value
- Footage of home (sqft_living)
- Footage of the lot (sqft_lot)
- No. of bathrooms
- Grade

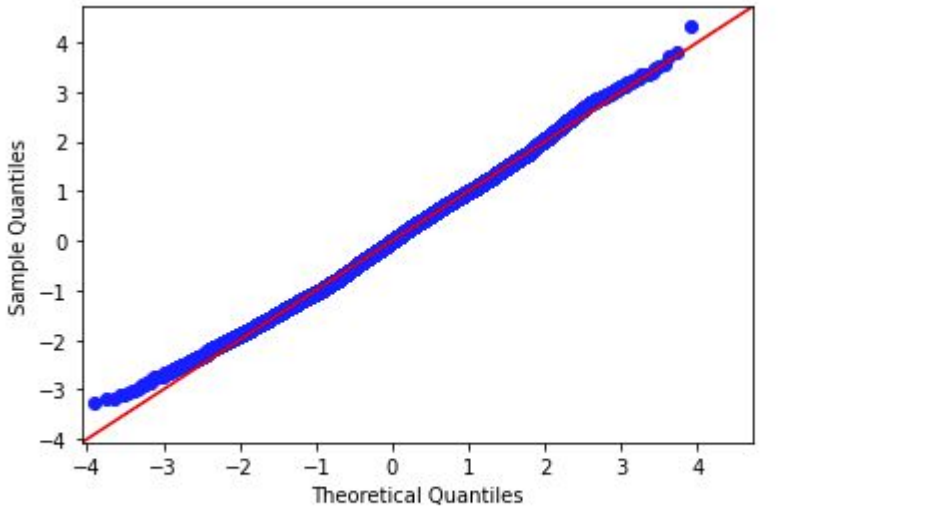
OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | price_log | **R-squared:** | 0.516 |
| **Model:** | OLS | **Adj. R-squared:** | 0.516 |
| **Method:** | Least Squares | **F-statistic:** | 5760. |
| **Date:** | Sun, 12 Feb 2023 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 17:50:02 | **Log-Likelihood:** | -22803. |
| **No. Observations:** | 21597 | **AIC:** | 4.562e+04 |
| **Df Residuals:** | 21592 | **BIC:** | 4.566e+04 |
| **Df Model:** | 4 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | 0.5617 | 0.013 | 44.398 | 0.000 | 0.537 | 0.587 |
| **sqft_living_log** | 0.5411 | 0.006 | 89.843 | 0.000 | 0.529 | 0.553 |
| **sqft_lot_log** | -0.1006 | 0.005 | -20.074 | 0.000 | -0.110 | -0.091 |
| **baths_5+baths** | 0.9087 | 0.107 | 8.487 | 0.000 | 0.699 | 1.119 |
| **quality_low** | -0.7016 | 0.015 | -48.041 | 0.000 | -0.730 | -0.673 |

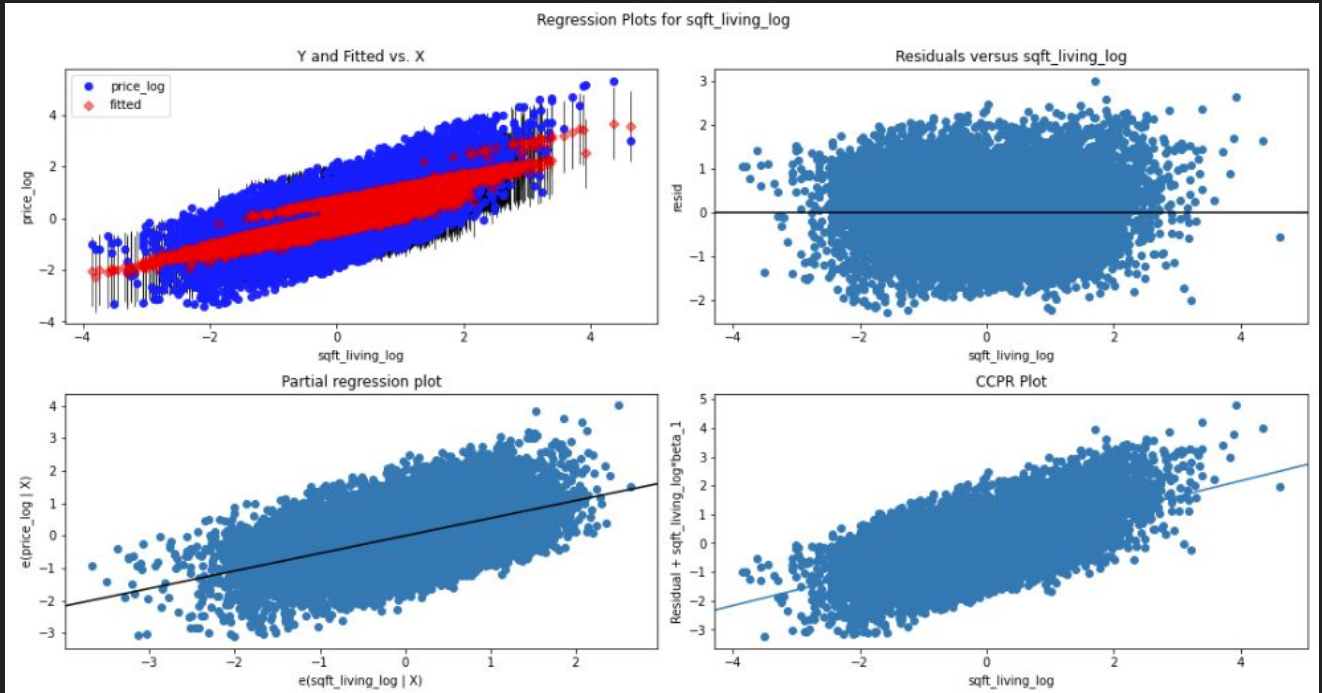
| | | | |
|---|---|---|---|
| **Omnibus:** | 67.763 | **Durbin-Watson:** | 1.975 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 65.605 |
| **Skew:** | 0.116 | **Prob(JB):** | 5.68e-15 |
| **Kurtosis:** | 2.862 | **Cond. No.** | 30.1 |

# Model Validation
# Normal Distribution assumption



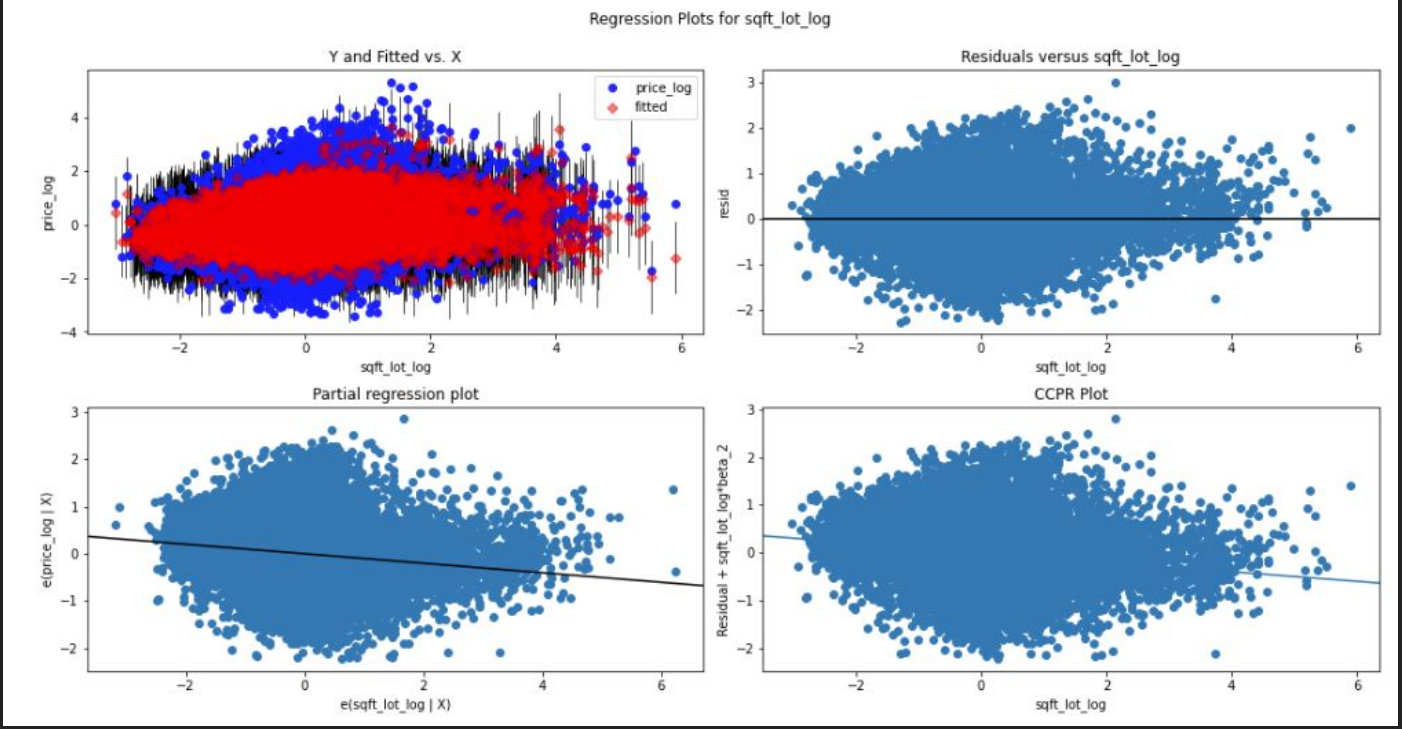

QQ plot shows that the normality assumption of the residuals seems fulfilled.

# Regression plot of the variables




From the first and second plot in the first row, we see the residuals appear to be equal across the regression line which is a sign of Homoscedasticity.

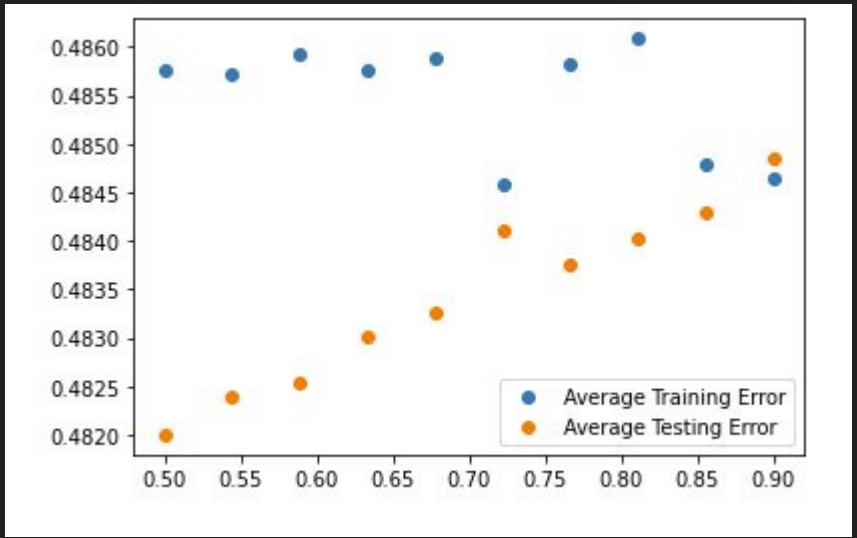
# Regression plot of the variables



# Train test

```
Train Mean Squared Error: 0.48344525331522825
Test Mean Squared Error:  0.48512251743393336
```

The difference between Test MSE and Train MSE is quite small (appx 0.35%)



# Thank you