Housing Sale Price prediction with the King County Housing Dataset

Pinal Patel

Academic Xi

Process

- The Data
- Data cleaning and pre-processing.
- Exploring the data
- Building regression models
- Validation of the model
- Result

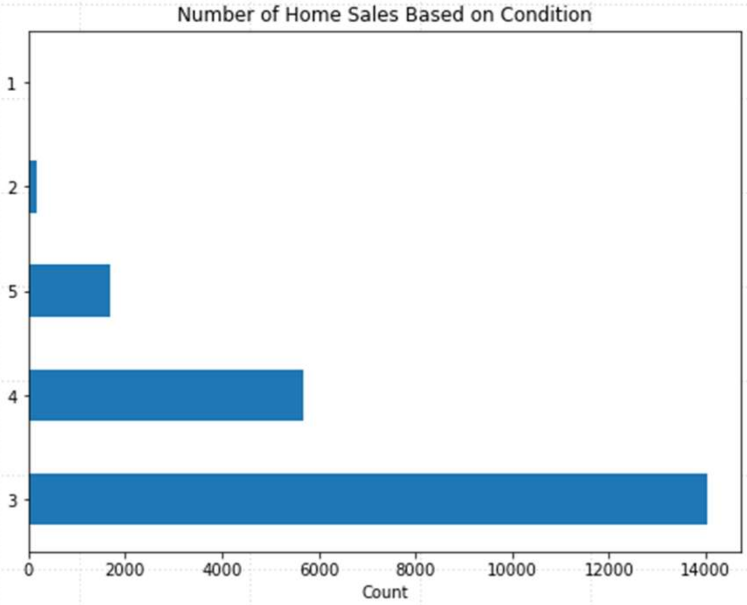
The Data

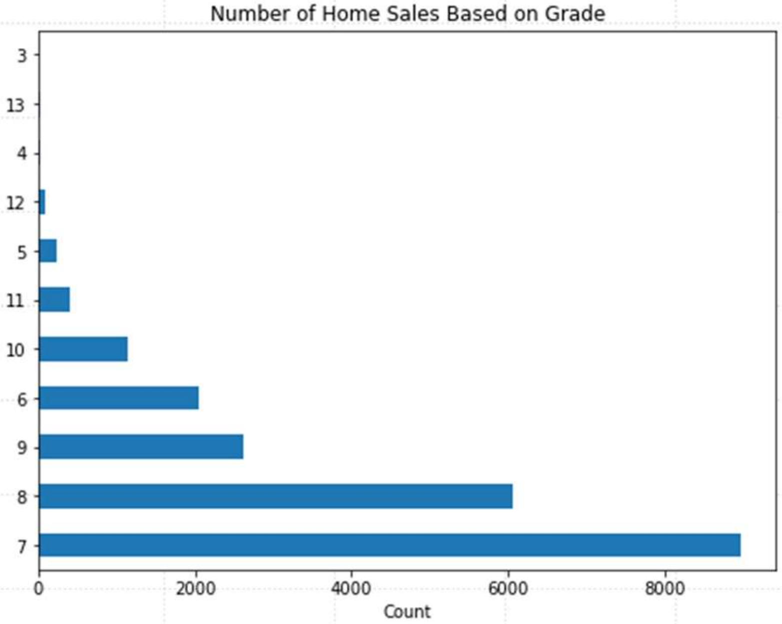
- The King County Housing Data Set contains information about the size, location, condition, and other features of houses
- Id, date, price bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, grade, sqft_above, sqft_basement, yr_built, yr_renovated, zipcode, lat, long sqft_living15, sqft_lot15

Business Problem

- RGB need to provide prospective home sellers with guidance on how to improve the value of their home prior to listing, including the predicted increase in value expected based on improvements to features.

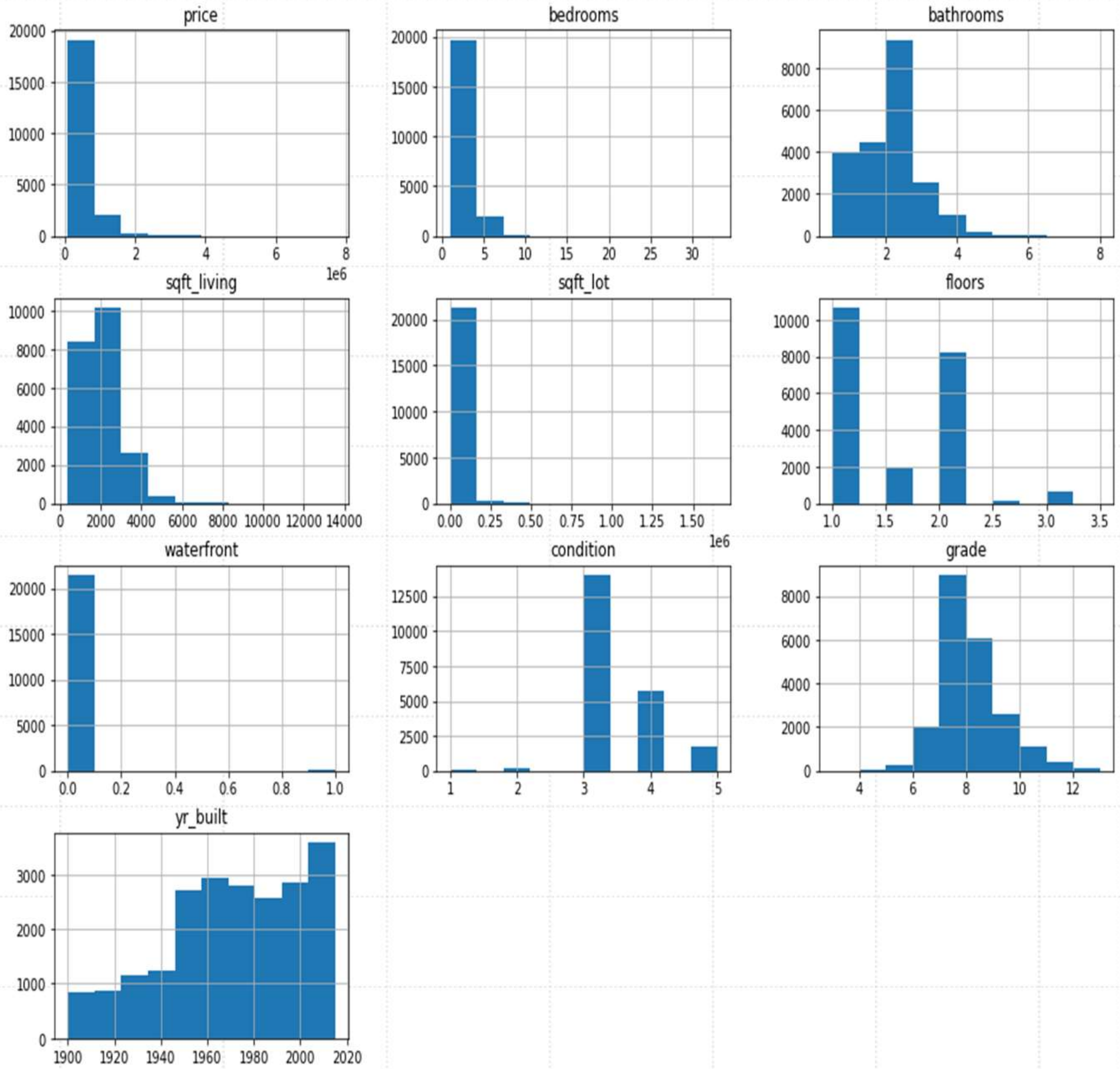
Exploring the Data



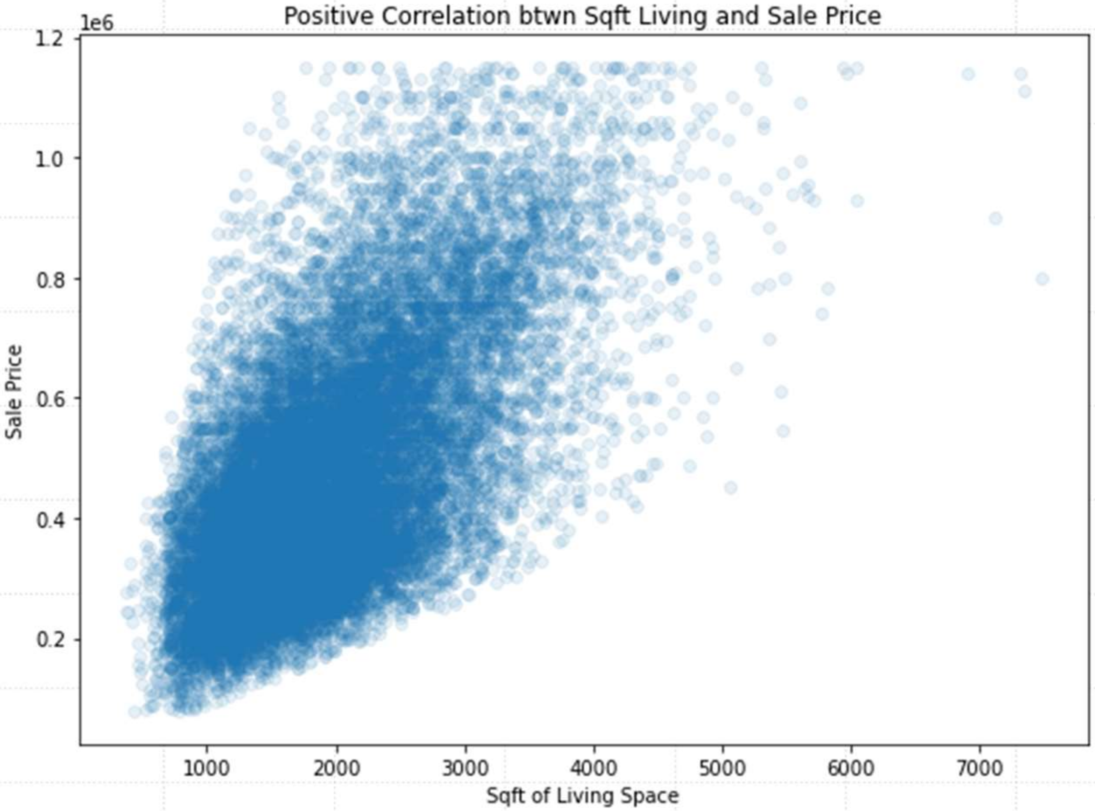


Histogram

- Many of the variables do not follow a normal distribution, and the scales are dramatically different for some variables. This may create issues with satisfying all regression assumptions, but we'll address those issues as they arise. Regression does not require features to be normally distributed

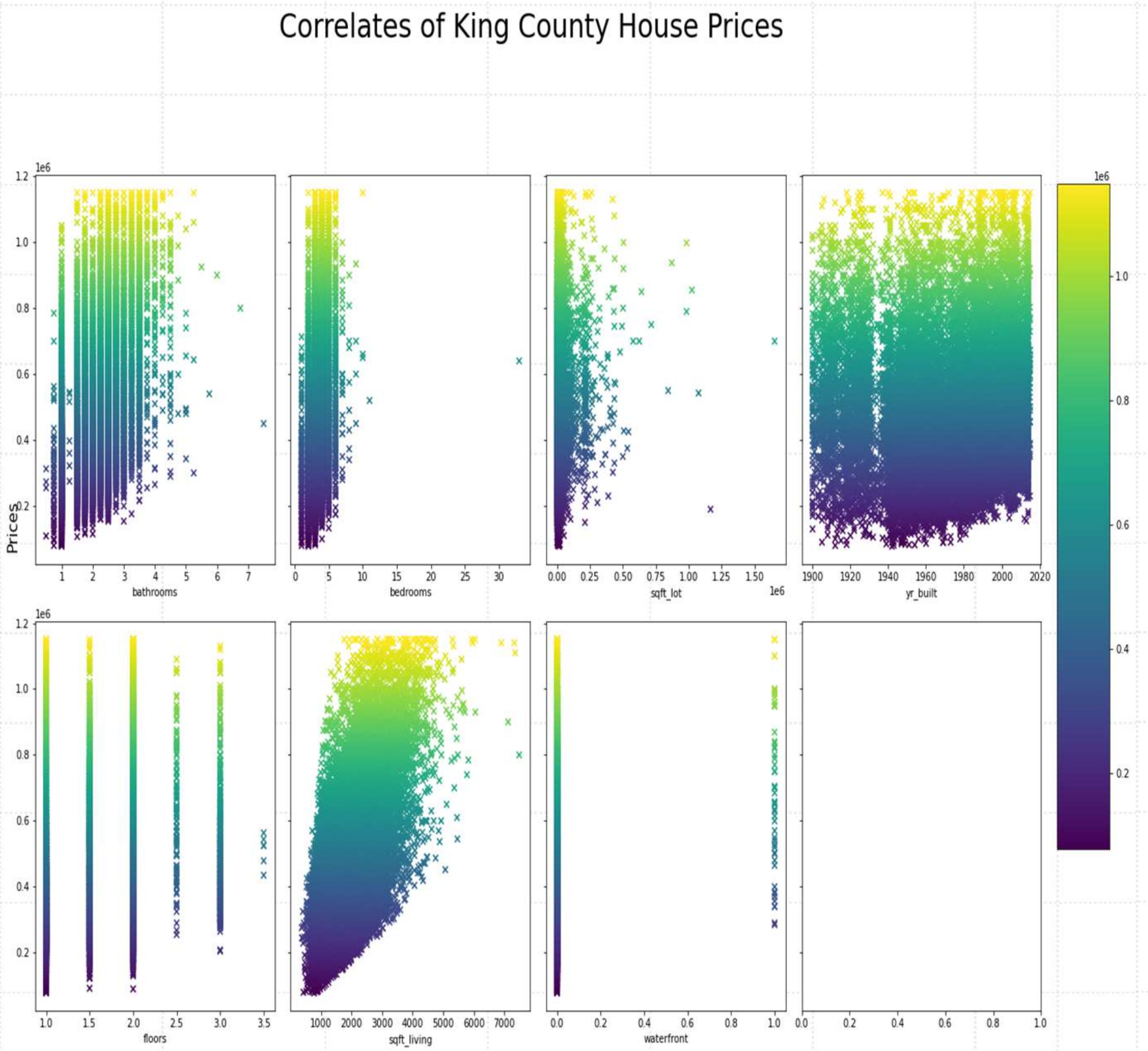


Checking correlations



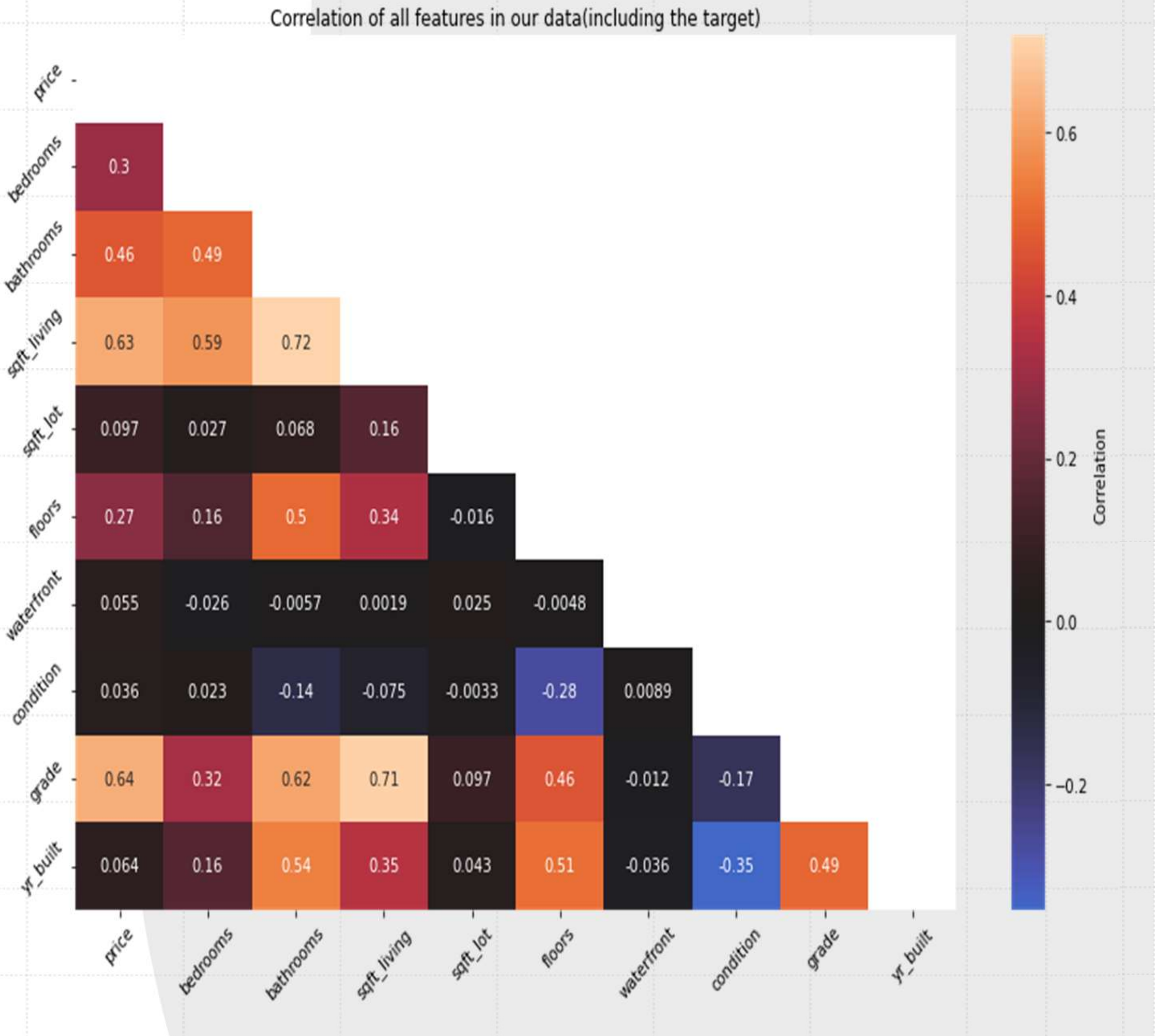
Checking correlations

- There is no obvious linear correlation between the variables "floors," and "bedrooms" and “waterfront” of a home. Sqft living will be employed in the multiple regression model because it has a better linear association with pricing than sqft lot.

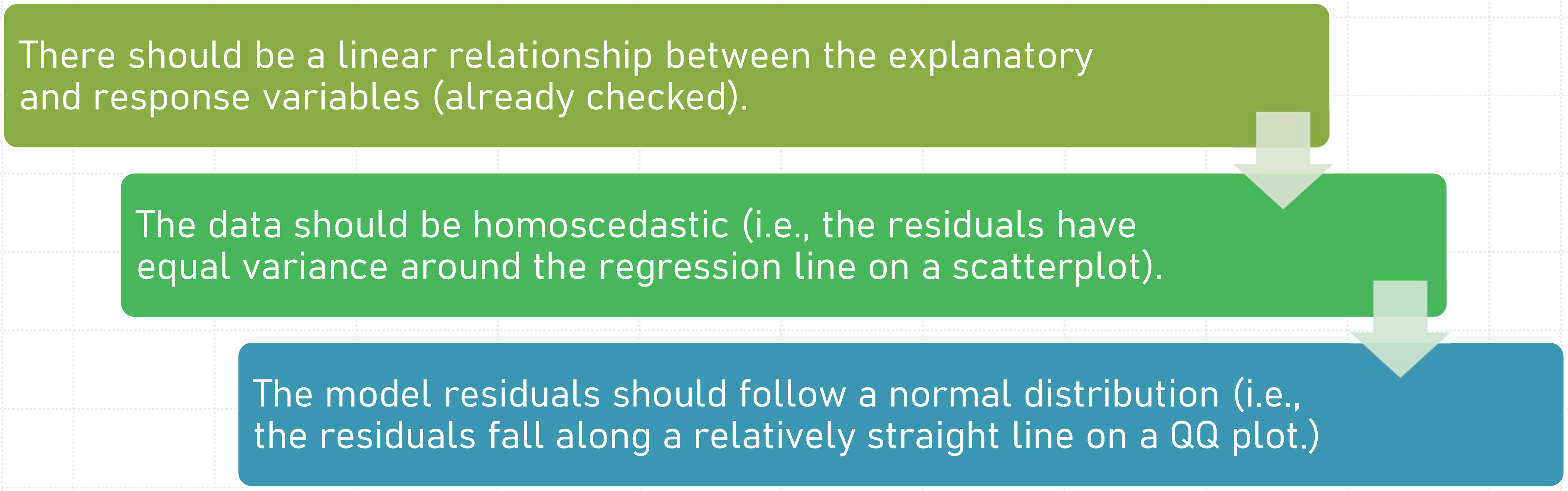


Multicollinearity

- price 1.000000
- grade 0.635934
- sqft_living 0.627438
- bathrooms 0.460266
- bedrooms 0.298955
- floors 0.273616
- sqft_lot 0.097296
- yr_built 0.063956
- waterfront 0.055232
- condition 0.035714
- Name: price, dtype: float64



Model Testing



There should be a linear relationship between the explanatory and response variables (already checked).

The data should be homoscedastic (i.e., the residuals have equal variance around the regression line on a scatterplot).

The model residuals should follow a normal distribution (i.e., the residuals fall along a relatively straight line on a QQ plot.)

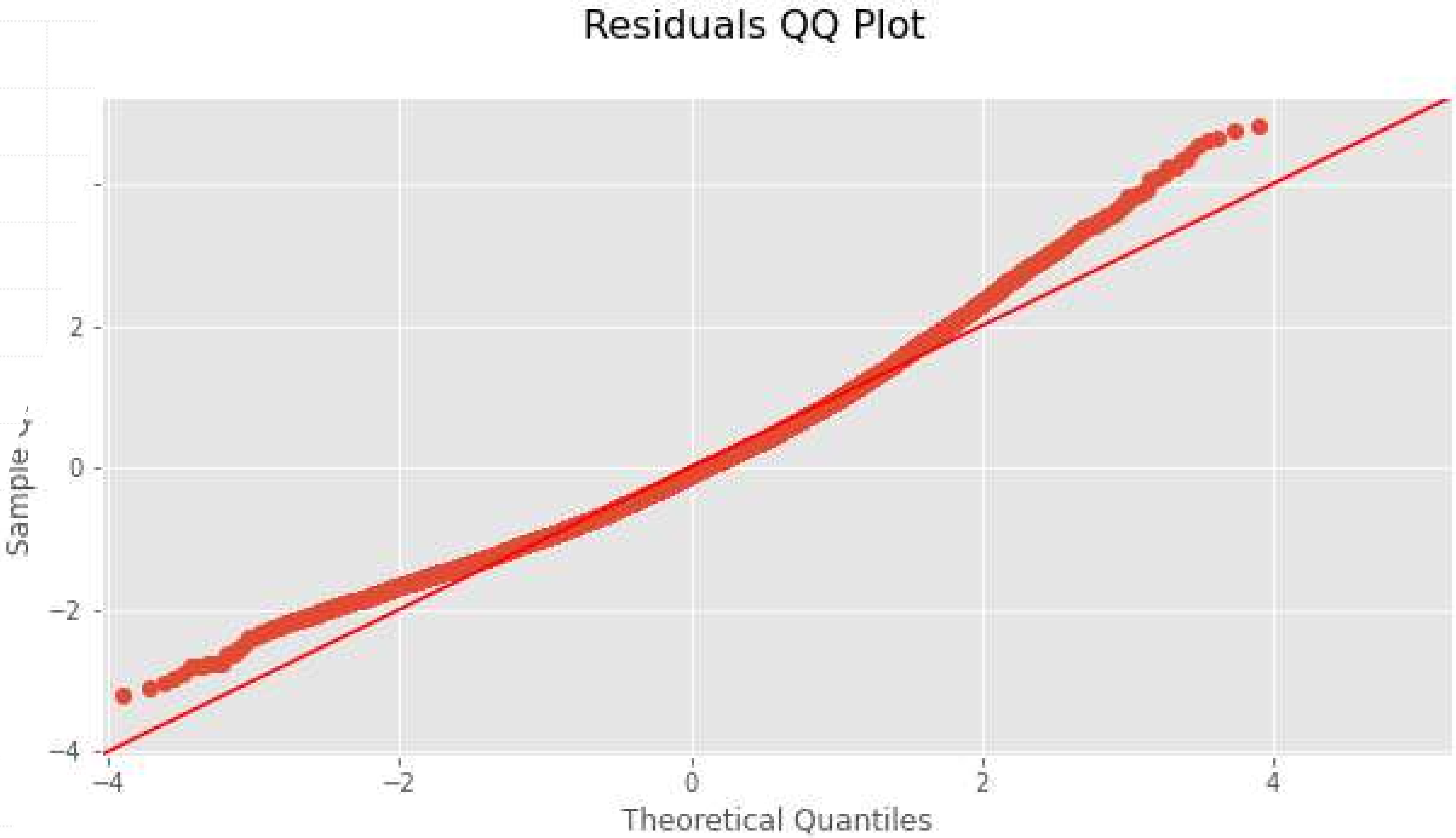
Train Mean Squared Error: 2.14597995431184e-18 Test Mean Squared Error: 2.1514436620943857e-18

Final Model and Results

Correlations   Features

- 0   0.718648 [bathrooms, sqft_living]
- 1   0.718648 [sqft_living, bathrooms]
- 2   0.709002 [sqft_living, grade]
- 3   0.709002 [grade, sqft_living]
- price ~ sqft_living + grade + bathrooms
- R 2=.47
- P < 0.05

- Train Mean Squared Error: 2.14597995431184e-18
- Test Mean Squared Error: 2.1514436620943857e-18


Residuals QQ Plot
Sample Q
−4
−2
0
2
−4
−2
0
2
4
Theoretical Quantiles

Recommendations

- Improve construction quality
- Expand living area
- Add a bathroom

Thank You

- https://github.com/pinalnikhil/KingCountymultilinearregression.git