NEW BUILDER KING COUNTY

Academy Xi

Phase 2 Project

Jules Mejia

Sunday 2 July 2023

Summary

- Successful residential builder
- Generated 5 key variables:
  - grade_11
  - grade_12
  - bathrooms_3.75
  - view_3.0
  - view_4.0
- Recommendation:
  - Custom design, high quality
  - 3 or more bathrooms
  - Location with an interesting view

Business Problem

Successful residential builder from the west coast of USA

King County, Washington

Variables strongly related to price

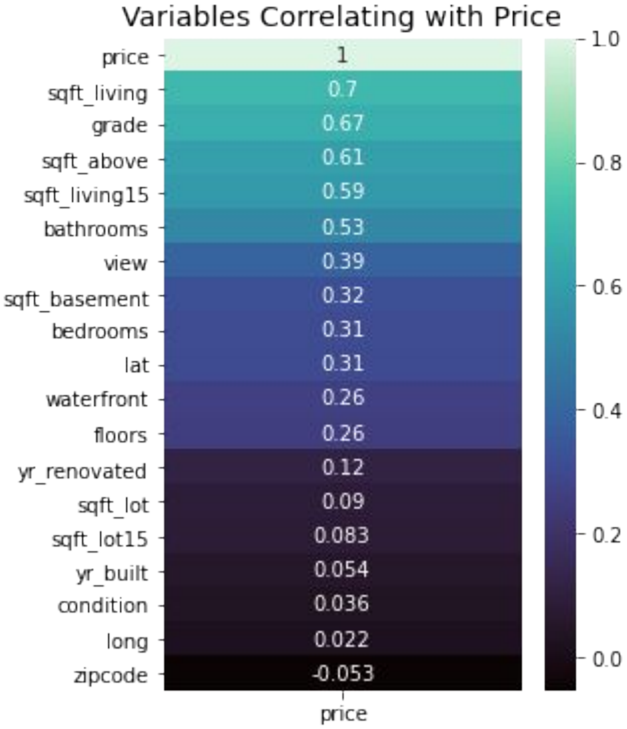
Data

Key Characteristics

- Houses sold in King County between 2014-2015
- 21,597 entries
- House features, location, ratings

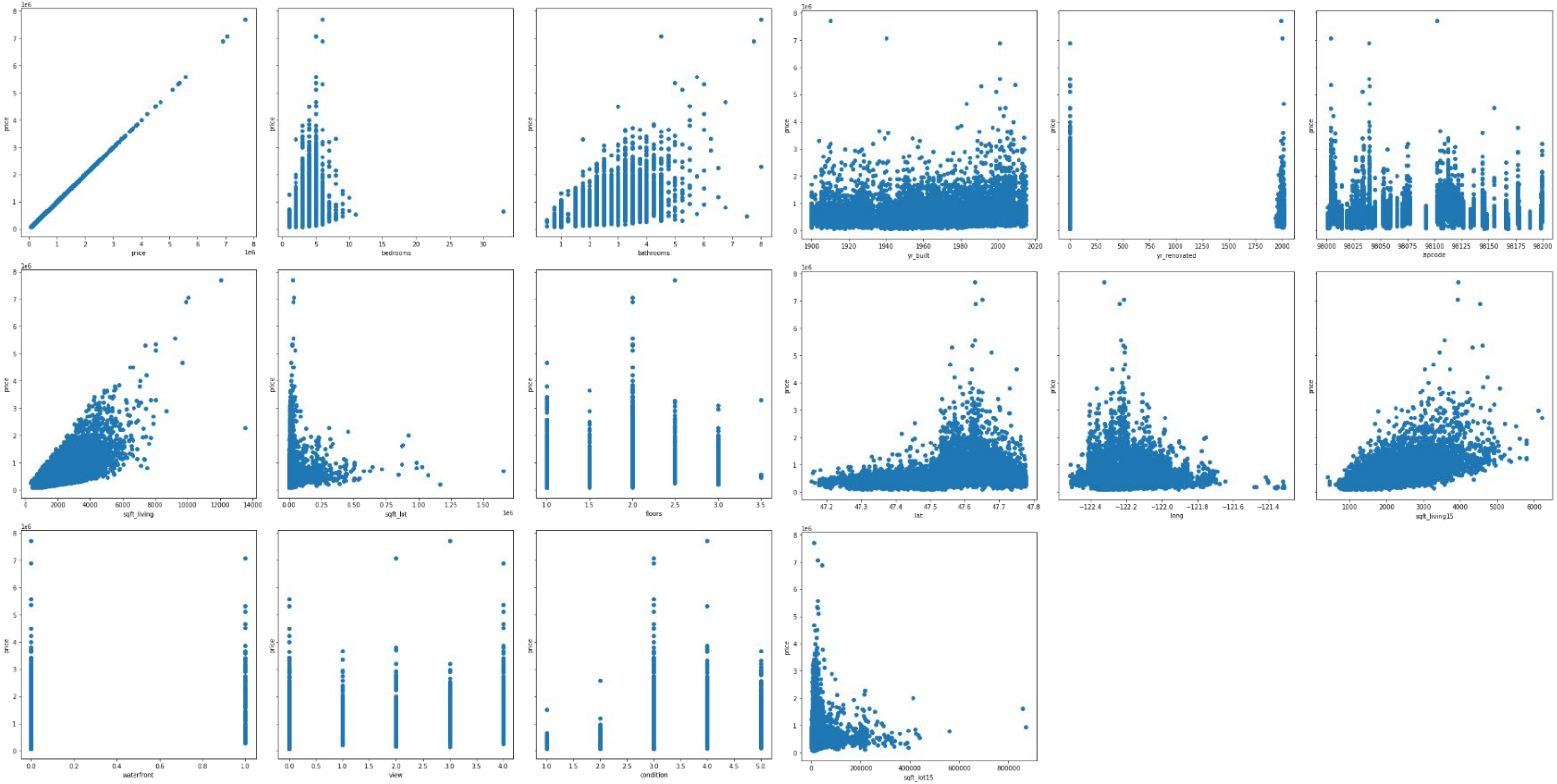
King County

Iteration 1 - Baseline Model



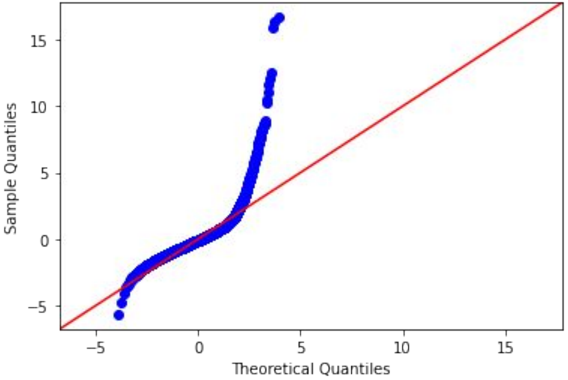
- Adjusted R-squared: 0.845
- Skew = 2.235 Positively skewed
- Kurtosis = 43.117 leptokurtic curve. Heavy tails, many outliers

Iteration 1 - Scatter Plot

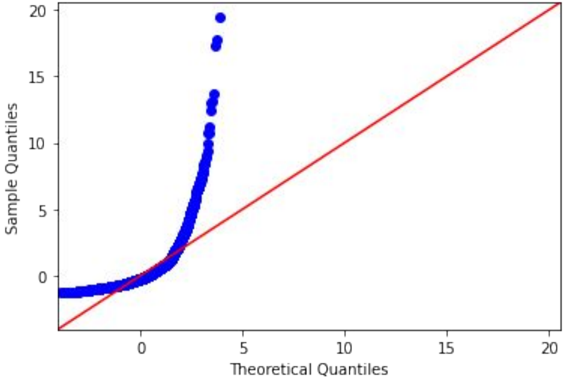


Iteration 1 - Q-Q Plots

sqft_living



long



Strong linear relationship w/ R-squared value

- sqft_living - 0.49
- sqft_above - 0.37
- sqft_living15 - 0.34

Possible strong linear relationship if there is a basement present

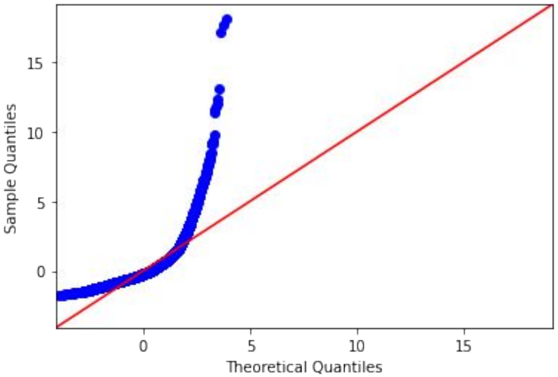
- sqft_basement - 0.10

Weak linear relationship

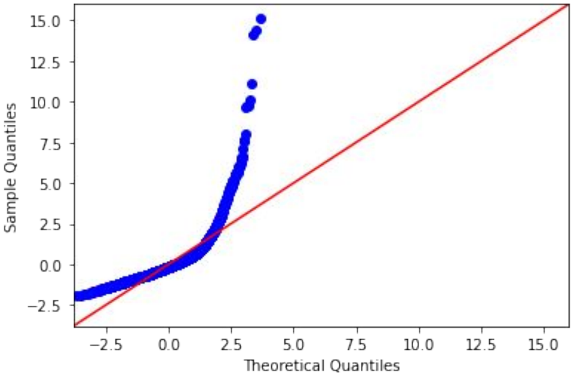
- sqft_lot - 0.008
- lat - 0.09
- long - 0.0004
- sqft_lot15 - 0.006

Iteration 1 - sqft_basement

Zeroes



No Zeroes



- R-squared with zeroes: 0.10
- R-squared without zeroes: 0.16

Iteration 2 - Dropping variables

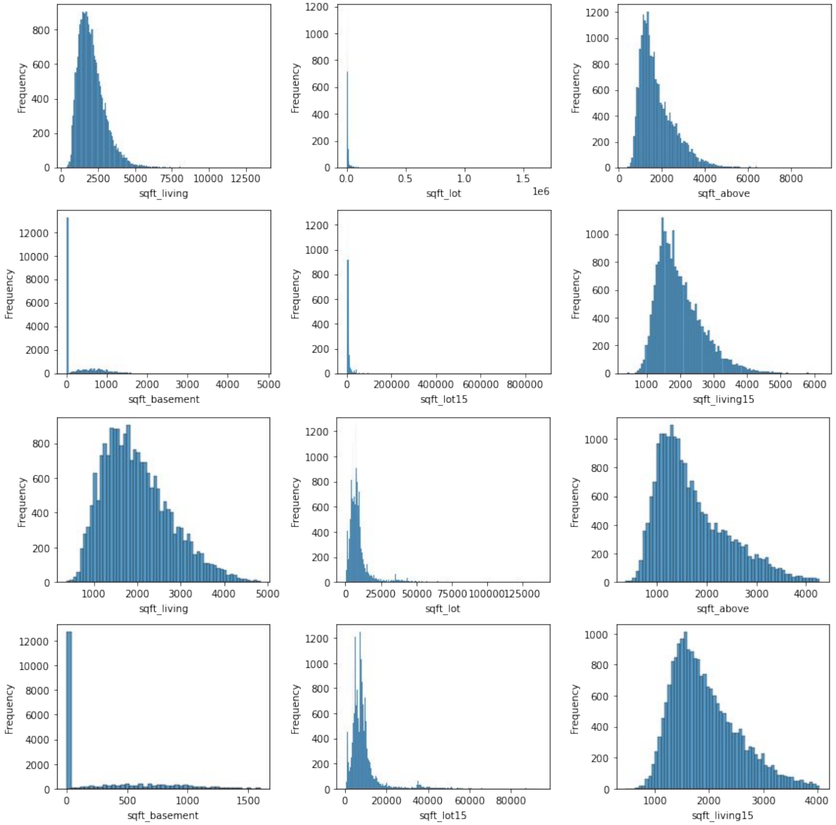
Categorical variables with correlation score

- grade - 0.67
- bathrooms - 0.53
- view - 0.39
- bedrooms - 0.31
- floors - 0.26
- waterfront - 0.26
- yr_renovated - 0.12
- yr_built - 0.054
- condition - 0.036
- zipcode - -0.053

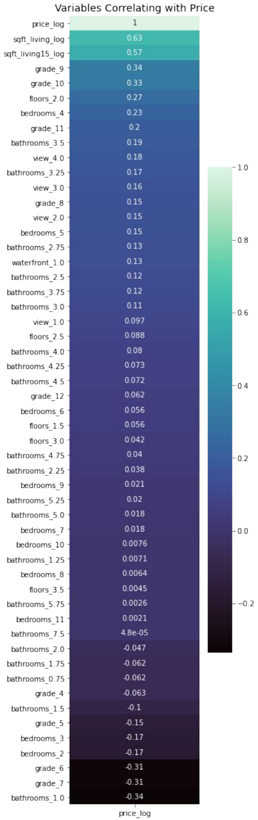
Weak linear relationship

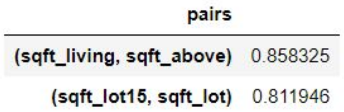
- sqft_lot
- lat
- long
- sqft_lot15
- sqft_basement

Iteration 2 - 3 standard deviations



Iteration 2 - Model

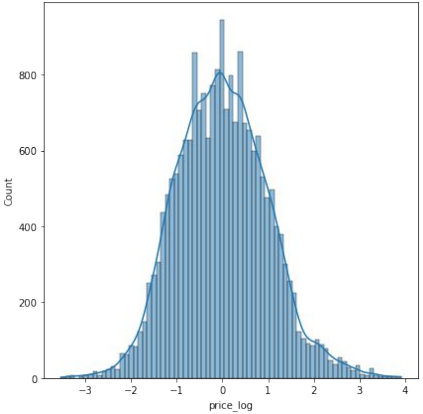


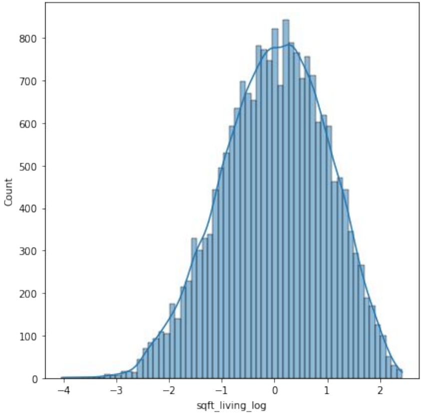


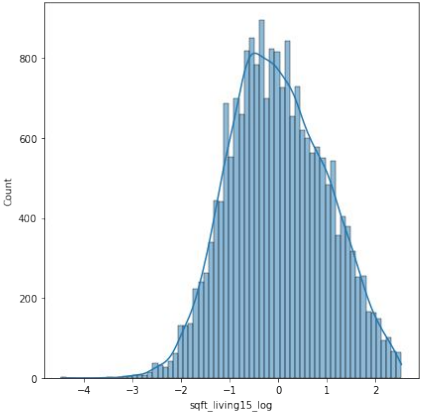
pairs	
(sqft_living, sqft_above)	0.858325
(sqft_lot15, sqft_lot)	0.811946

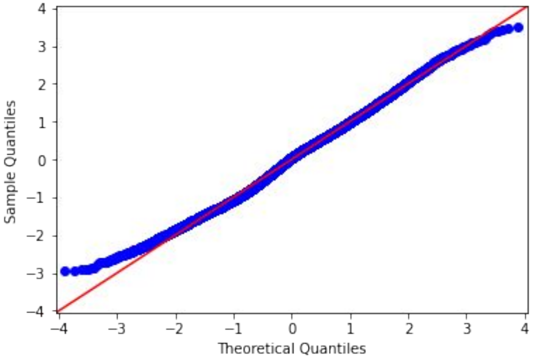
- Adjusted R-squared: 0.558
- Skew = -0.023 No skew
- Kurtosis = 2.293 mesokurtic

Iteration 2 - Assumptions









sqft_living

Iteration 3 - Interactions and Model Validation

Categorical variables with correlation score

- 'sqft_living_log', 'grade_4', 0.555
- 'sqft_living_log', 'bedrooms_2', 0.555
- 'sqft_living_log', 'waterfront_1.0', 0.555

Adjusted R-squared: 0.557

Train Test Split

- Train Mean Squared Error: 0.4426173699654987
- Test Mean Squared Error: 1.7332208309125228e+20

Overfitting

Iteration 4 - Reduce Variance

- grade_4, grade_5, grade_6, grade_7, grade_8, grade_9, grade_10
- bathrooms_0.75, bathrooms_1.0, bathrooms_1.25, bathrooms_1.5, bathrooms_1.75, bathrooms_2.0, bathrooms_2.25, bathrooms_2.5, bathrooms_2.75, bathrooms_3.0, bathrooms_5.0, bathrooms_5.75, bathrooms_7.5, bathrooms_6.0
- bedrooms_8, bedrooms_9, bedrooms_10, bedrooms_11
- Floors_3.5

2 standard deviations

Iteration 4 - Model Validation

- Train Mean Squared Error: 0.5740418030948392
- Test Mean Squared Error: 0.5772426558303619

Evaluation

Top 5 variables with a strong relationship to price

- grade_11
- grade_12
- bathrooms_3.75
- view_3.0
- view_4.0

Conclusion

- Creating a custom design using high quality materials, high quality finish work and luxurious options
- Incorporating 3 or more bathrooms into their designs
- Choosing a location of the house with a great view of local points of interest

Thank You!

Email: julespmejia@gmail.com
GitHub: github.com/julesmejia
LinkedIn: linkedin.com/in/julesmejia/