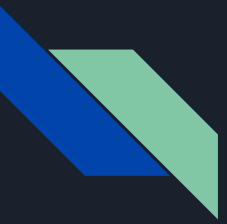
King County Housing Prices

Linear Regression Project

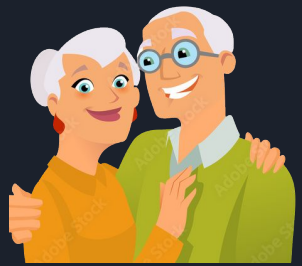
By Mike Van Eaton | Flat Iron School



Realtor

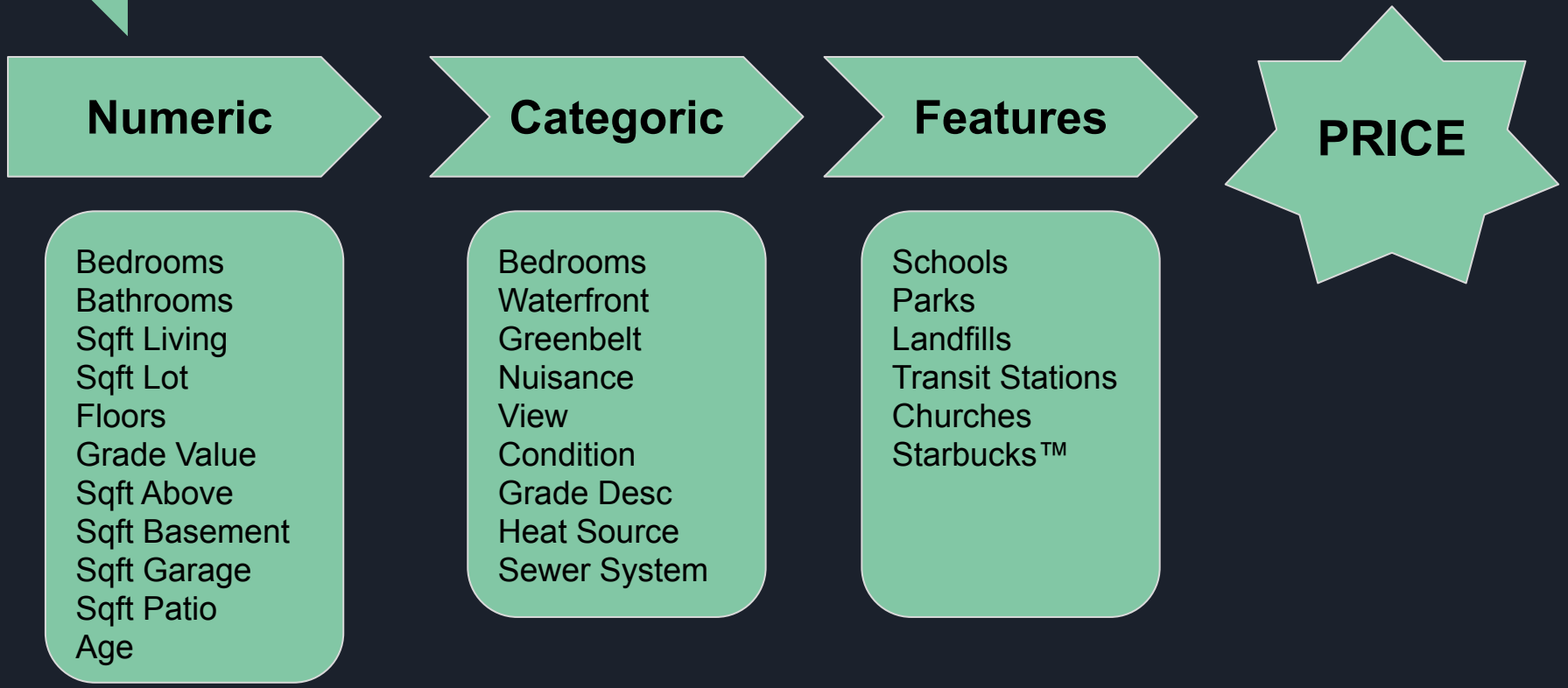
Appraisers

Home Buyers



Home Sellers

House Value Features



Model 1B

Top 5 Correlated Features

Sqft Liiving	0.64
Grade Value	0.62
Sqft Above	0.56
Bathrooms	0.50
Bedrooms	0.34

- Top Correlated Feature Used To Predict House Value Variance
- Price Outliers Removed

- R-sqrd: 41%
- Const: $150,000 for no house
- p-value: significant

Model 2A

- All Numeric Features used
- Price Outliers Removed

Bedrooms
Bathrooms
Sqft living
Sqft lot
Floors
Grade val
Sqft above
Sqft basement
Sqft garage
Sqft patio
Age

- R-sqrd: 51%
- Const: - $1,420,000 no features
- p-value: significant

Model 3

Waterfront Greenbelt
Nuisance
View Condition
Grade Desc Heat
Source
Sewer System Zip
Code

- All Categoric & Numeric Features used
- Price Outliers Removed

- R-sqrd: 75%
- Coef: $390,000 no features
- p-value: most are significant

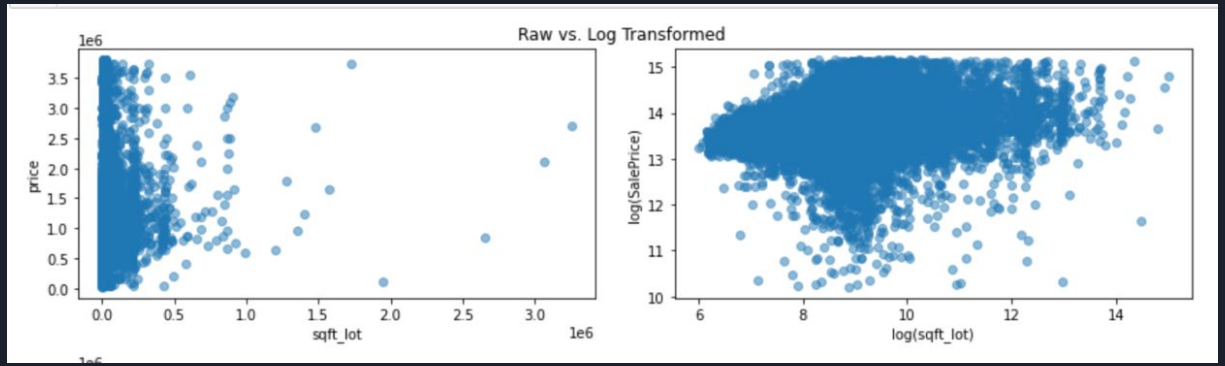
Model 4

- All Categoric , Numeric & Local Features used
- Price Outliers Removed

Public Schools
Parks
Landfills
Parks
Churches
Transit Stations
Starbucks™

- R-sqrd: 75%
- Coef: $400,000 no features
- p-value: most are significant

LINE Tests
Linearity

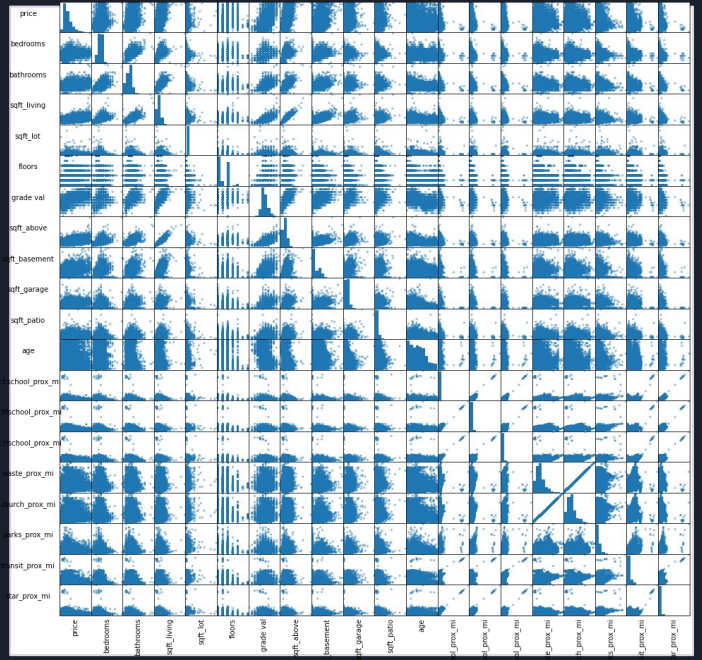


Sqft Lot
Sqft Patio
Bathrooms
Grade Val
Age
Sqft Living

Log Transformation for certain features that may be nonlinear

R-values compared.
Sqft Lot had a 2% increase with the transformation

LINE Tests
Independence

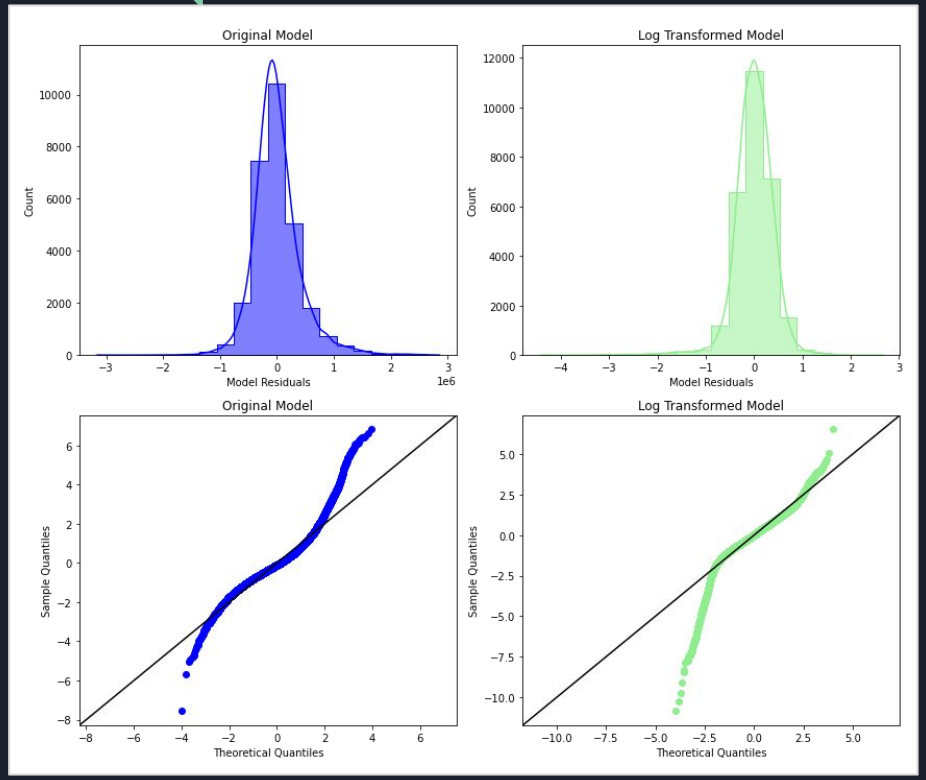


All features compared for independence

Threshold: ≥ 0.75 correlation

Collinear Pairs	
Sqft Living, Sqft Above	0.87
Elementary Sch, Starbucks™	0.78
Sqft Living, Bathrooms	0.76

LINE Tests
Normality



Compare residual distribution to Log transformed

QQ test

- **Leptokurtic - high Kurtosis**
- **Normal but Narrow Peak**

LINE Tests
Equal Variance

The Goldfeld-Quandt Test for Equal Variance was used on the

Sqft living feature.

A p-value greater than 0.05 is expected for consistent variance in the residuals

p- value for sqft living is 0.

The Goldfeld_Quandt test between price and sqft_living shows that there is inconsistent variance in the residuals.

Final Model

- All Categoric , Numeric & Local Features used
- Features Removed
  - Price Outliers
  - Sqft Lot
  - Starbucks™
  - Sqft Above
  - Bathrooms
- R-sqrd: 74%
- Coef: $410,000 no features
- p-value: most are significant

Conclusion

	Model 1b	Final Model
R-squared	41%	74%
Const	$150,000	$410,000
Features	1	21

Next Steps:
Reduce the number of features
Reduce kurtotic effects
Increase R-squared predictive value