# Predicting House Prices with the King County Housing Dataset

Avonlea Fisher
September 2020

# The Data

The King County Housing Data Set contains information about the size, location, condition, and other features of houses in Washington's King County.

After data cleaning:

- 21,000 house sales
- 20 variables

# Process

01 Data cleaning and preprocessing

02 Exploring the data

03 Building linear models

04 Interpreting model results

# Correlations

Correlations with Price

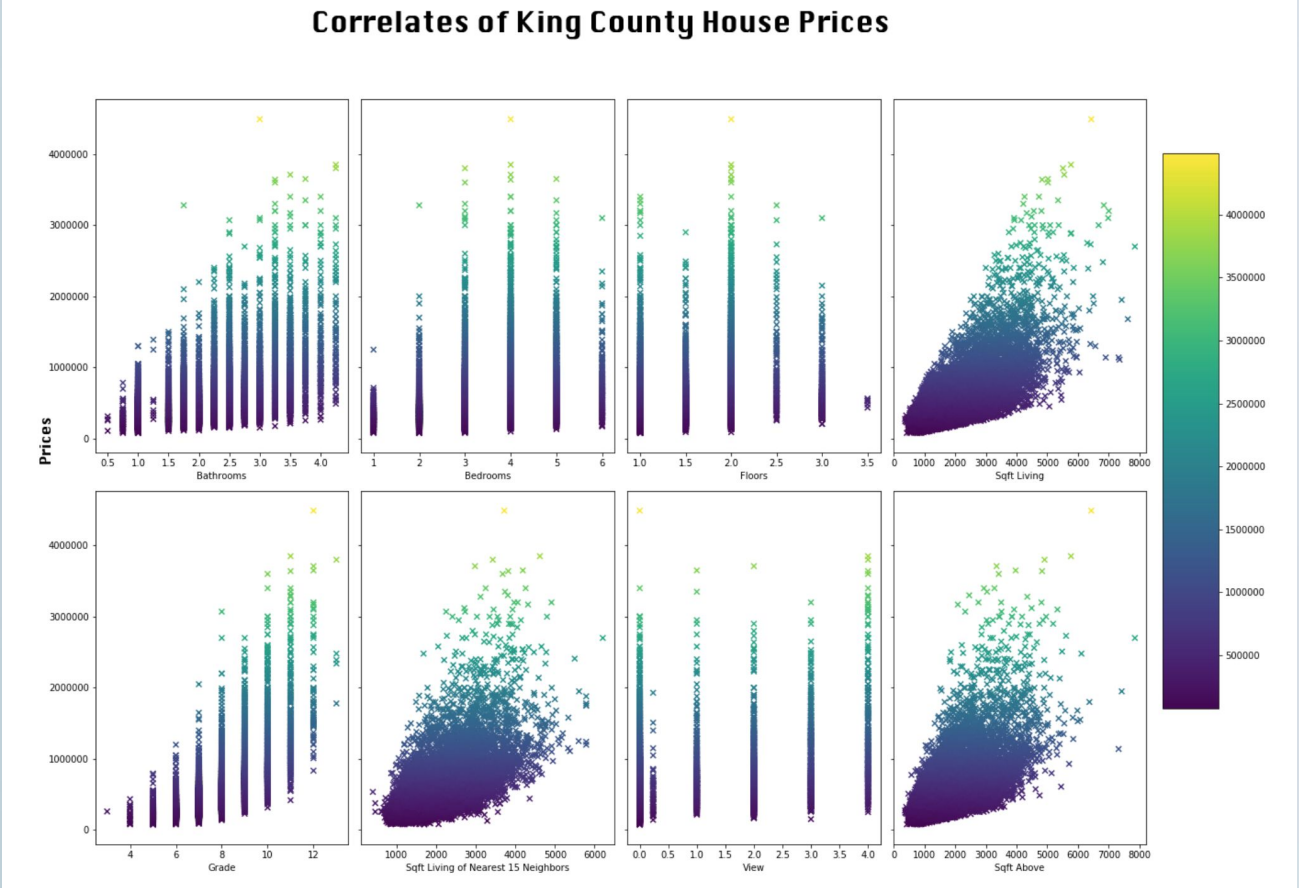
| | Correlations | Features |
|---|---|---|
| 2 | 0.677596 | sqft_living |
| 4 | 0.668335 | grade |
| 6 | 0.593674 | sqft_living15 |
| 5 | 0.578363 | sqft_above |
| 1 | 0.489138 | bathrooms |
| 3 | 0.386794 | view |
| 0 | 0.302105 | bedrooms |

## Notes

-Features that were highly correlated with price were considered for inclusion in the model.

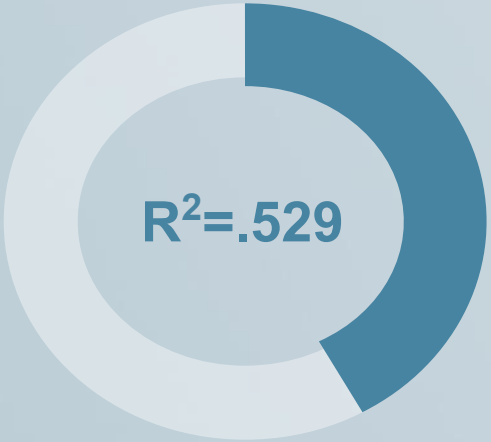
# Notes

- View, floors, and bedrooms were excluded from models due to a weak linear relationship



# Final Model and Results

price ~ sqft_living + grade + bathrooms

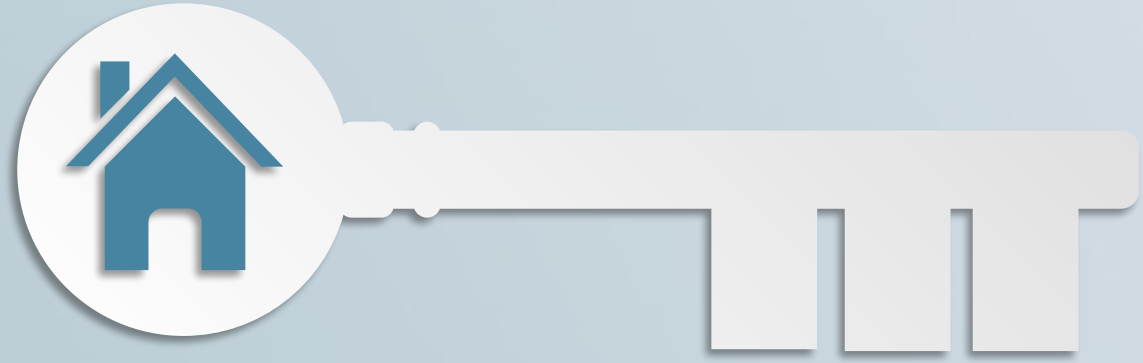




**Notes: The model can account for about 53% of the variability in price. A p-value of less than 0.05 means that we can reject the hypothesis that there is no relationship between price and the predictor variables.**

# Recommendations

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



The square footage of a house, its grade, and number of bathrooms are among the strongest predictors of house prices.

# Questions for Future Work

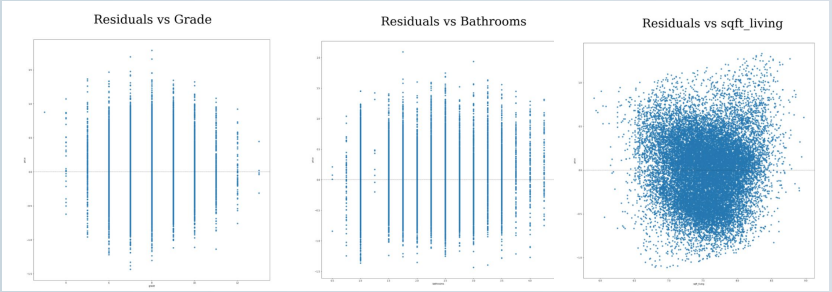
- Are the best predictors of price similar for homes outside of King County?
- How does the analysis change if no extreme values are excluded from the dataset?

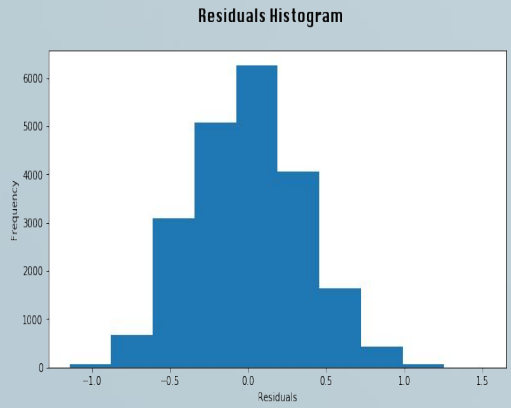
# Thank you!

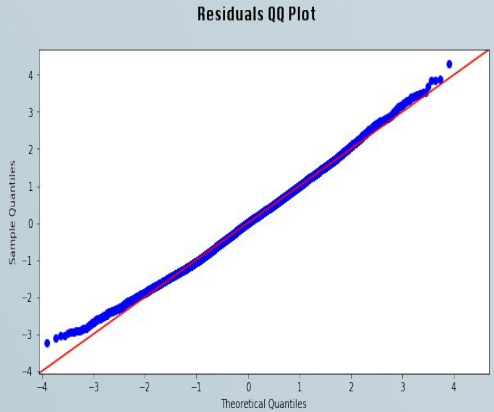
# Appendix

# Final Assumption Checks

**Homoscedasticity:** the residuals for all predictors have mostly equal variance along the regression line.




**Normality:** the residuals follow a normal distribution.

# Multicollinearity

Multicollinear Features

| | Correlations | Features |
|---|---|---|
| 0 | 0.866887 | [sqft_living, sqft_above] |
| 1 | 0.866887 | [sqft_above, sqft_living] |
| 2 | 0.812117 | [3, 4] |
| 3 | 0.812117 | [4, 3] |

**Notes**

-Combinations of highly multicollinear features were avoided in the models.