# King County Housing Prices

MODULE 1 FINAL PROJECT

KRYSTIAN DENNIS 12 AUGUST 2019

# Contact Information

- Krystian Dennis
- krystiandennis@gmail.com
- Jupyter Notebook and Presentation can be found at:
- https://github.com/krystiandennis/mod1-final-project

# Outline

- 1 - Objectives and Dataset
- 2 - Initial Observations
- 3 - Cleaning the Data
- 4 - Preparing Data for Analysis
- 5 - Data Model
- 6 - Model Results
- 7 - Recommendations

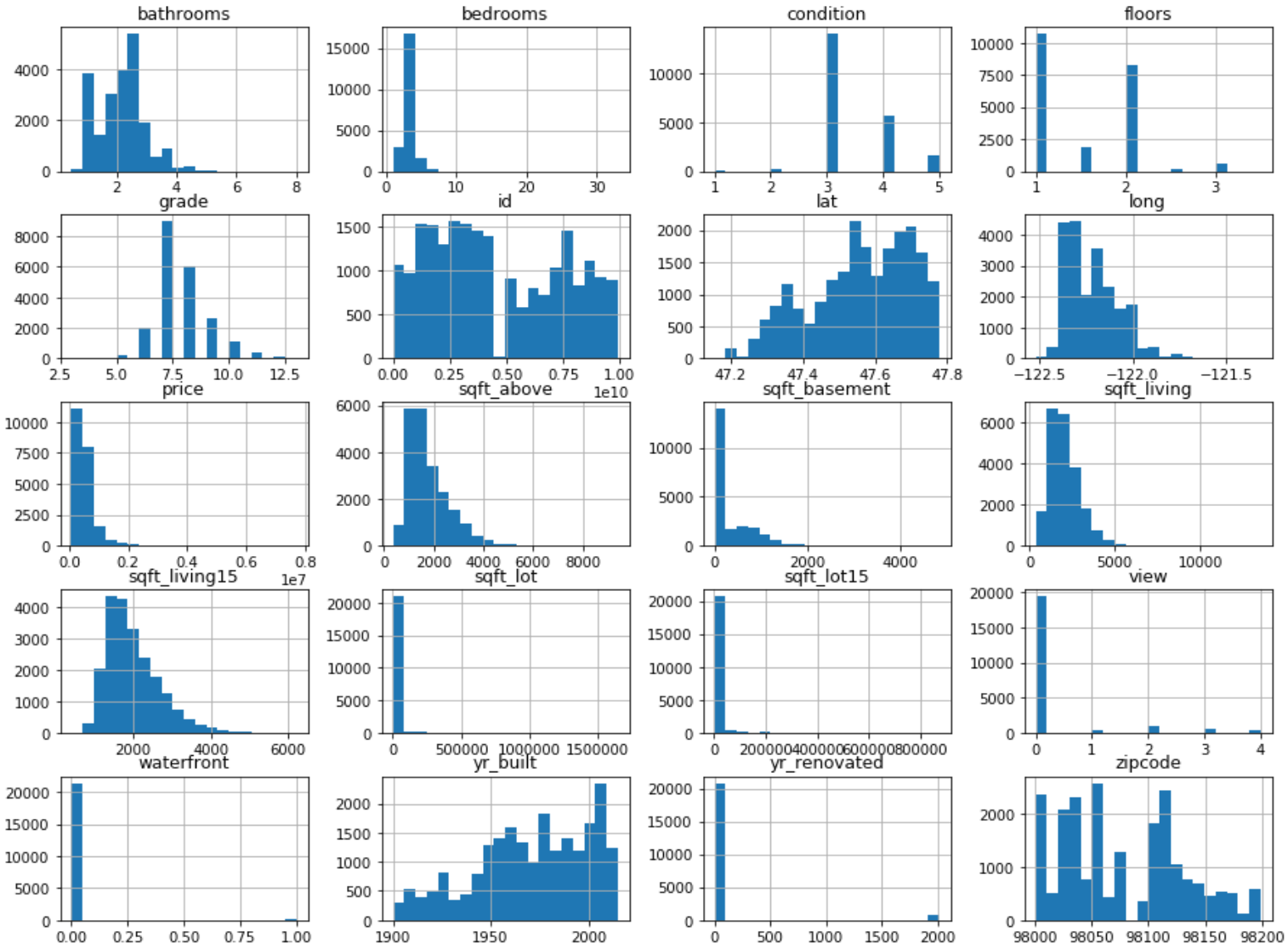
# Objective and Data Set

- The King County House Price Dataset contained records of over 21,000 home sales the Seattle area from Sept 2014-Sept 2015
- The dataset included 19 variables ranging from number of bathrooms to square footage of the basement
- The objective was to use these variables to create a model capable of predicting the sales price of similar properties in the area

| | | | | | | |
|---|---|---|---|---|---|---|
| Unique ID | Date sold | House price | Number of bedrooms | Number of bathrooms | Square footage (home) | Square footage (lot) |
| Number of floors | Waterfront view | Viewed or not | Overall condition | Overall grade by KC | Square footage (upper) | Square footage (basement) |
| Year built | Year renovated | Zip code | Latitude coordinates | Longitude coordinates | Square footage of living space (neighbors) | Square footage of lot (neighbors) |

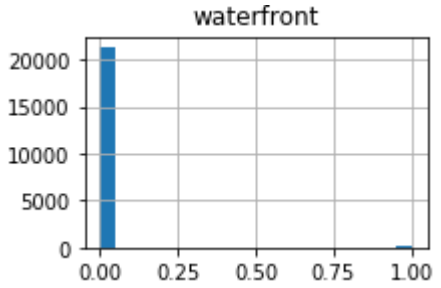
# Initial Observations

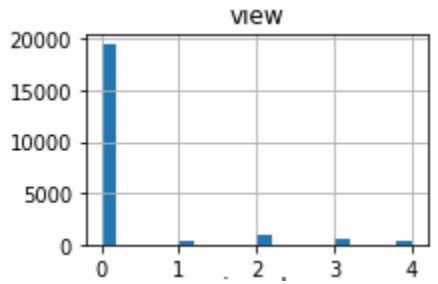
- The dataset contains categorical (like bedrooms) and continuous variable (like sqft_lot)
- The data appears skewed for variables like sqft_above
- There appears to be missing data for variables like yr_renovated

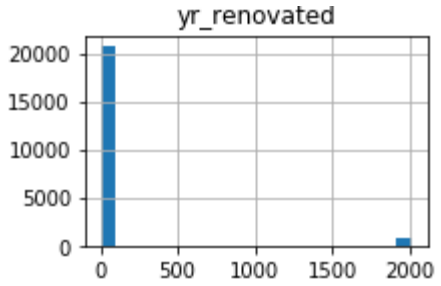


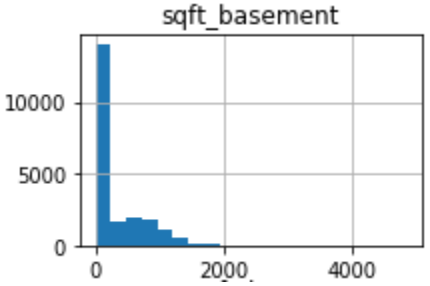
# Cleaning The Data

- Three variables were missing values
  - ‘waterfront’ was missing 11% of values
    - Replaced with ‘0’
  - ‘view’ was missing 0.2% of values
    - Replaced with ‘0’
  - ‘yr_renovated missing 18% of values
    - Replaced with ‘0’
- One variables contained placeholders
  - ‘sqft_basement’ contained 2% of ‘?’
    - Replaced with ‘0’



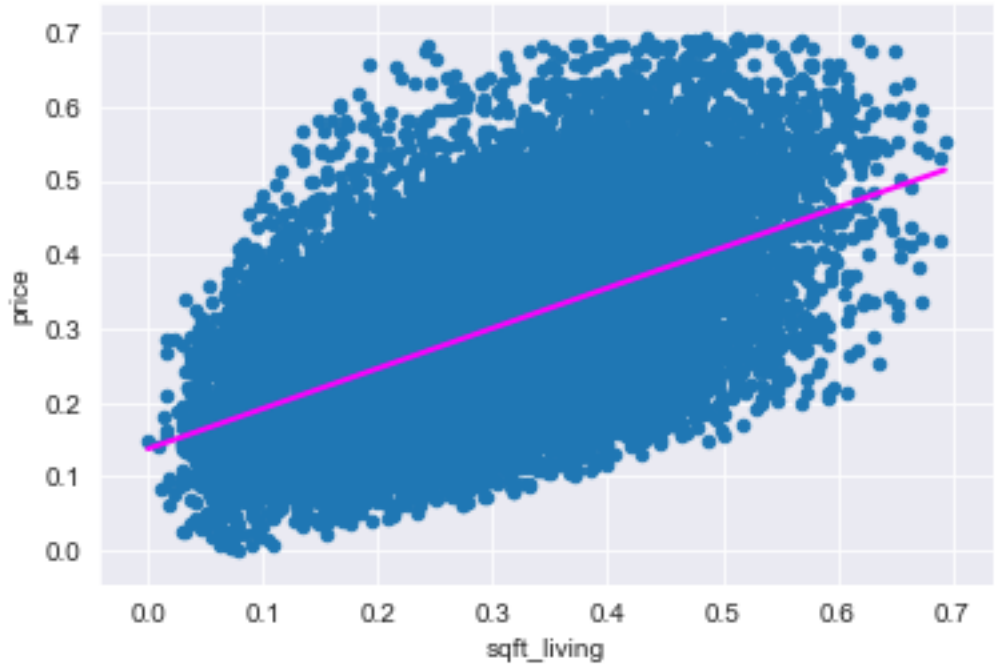
# Preparing Data for Analysis

- Variables not demonstrating linear relationship with 'price' removed from dataset
- Outliers removed from dataset
  - IQR*1.5
- All variables min-max scaled
- Continuous variables log transformed
- Independent variables with high correlation to each other removed from dataset

# Data Model

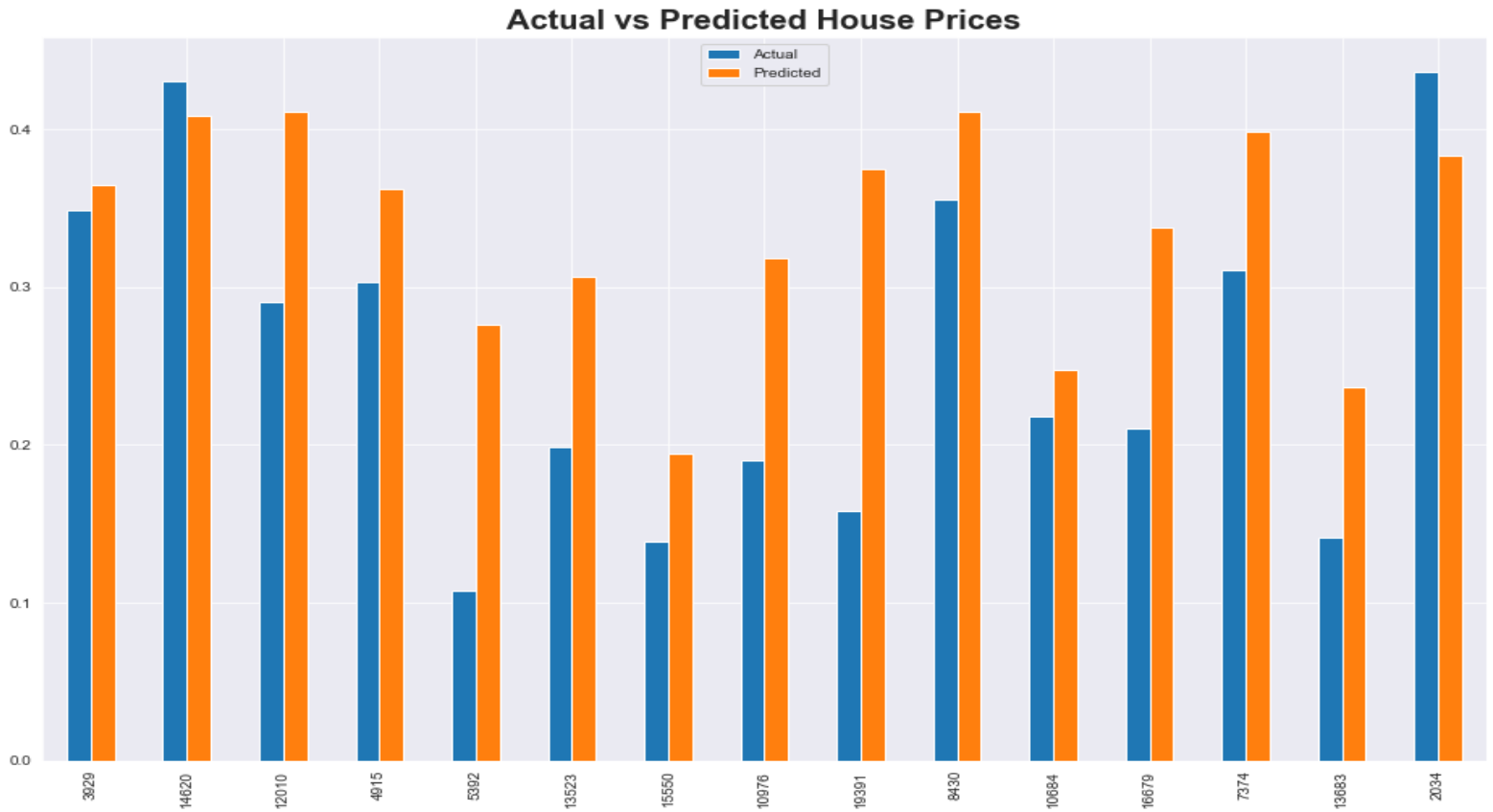
- Simple Linear Regression used on continuous variable to determine inclusion in model
  - r-squared above 0.2
  - p-value below 0.05
- Recursive Feature Elimination used to construct model
  - 2 variables
  - 4 variables
  - All variables

- Performance of model evaluated using linear regression



# Model Results

- Approximately 60% of variation in house prices can be explained by the model
- Model cross-validated with only -0.03 difference of MSE between actual and predicted house prices.




| | | | |
|---|---|---|---|
| **Dep. Variable:** | price | **R-squared:** | 0.597 |
| **Model:** | OLS | **Adj. R-squared:** | 0.597 |
| **Method:** | Least Squares | **F-statistic:** | 3901. |
| **Date:** | Sun, 11 Aug 2019 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 23:49:26 | **Log-Likelihood:** | 14267. |
| **No. Observations:** | 13154 | **AIC:** | -2.852e+04 |
| **Df Residuals:** | 13148 | **BIC:** | -2.848e+04 |
| **Df Model:** | 5 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | -0.0809 | 0.003 | -26.817 | 0.000 | -0.087 | -0.075 |
| **sqft_living** | 0.3380 | 0.010 | 34.692 | 0.000 | 0.319 | 0.357 |
| **bathrooms** | -0.0267 | 0.004 | -6.224 | 0.000 | -0.035 | -0.018 |
| **grade** | 0.1276 | 0.004 | 34.746 | 0.000 | 0.120 | 0.135 |
| **lat** | 0.4185 | 0.005 | 89.909 | 0.000 | 0.409 | 0.428 |
| **sqft_living15** | 0.0777 | 0.006 | 12.518 | 0.000 | 0.066 | 0.090 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 811.922 | **Durbin-Watson:** | 1.988 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 1049.265 |
| **Skew:** | 0.582 | **Prob(JB):** | 1.43e-228 |
| **Kurtosis:** | 3.747 | **Cond. No.** | 20.6 |

# Conclusion

- Up 60% of the variation in house prices can be explained by the model
- ‘sqft_living’ and ‘lat’ influence house prices most strongly
- Model tends to predict a higher house price than the actual house prices
  - To improve model, more variables will be included in model