King County House Sales Study

Flatiron School

Kamile Yagci 12/14/2021

Content

- Overview
- Business Problem
- Method
- Data
- Model
- Interpret
- Future Work
- Questions

Overview

- This project is the analysis of the King County House Sales.
- The Multiple Linear Regression is used to model the data and predict the house sale price.

Project Details:

- GitHub: https://github.com/kamileyagci/dsc-phase-2-project

Business Problem

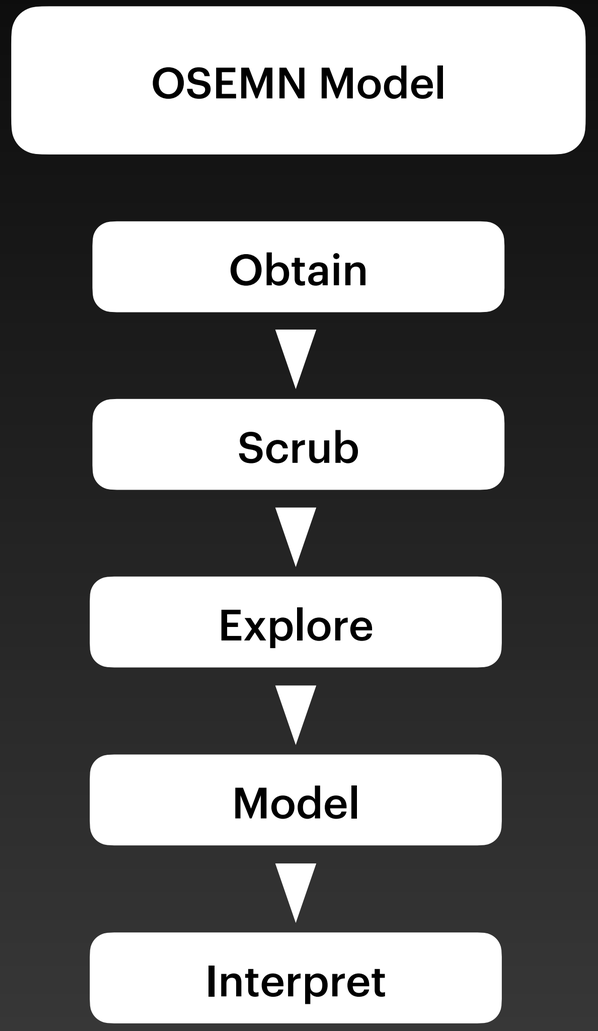
The Windermere Real Estate Agency request an analysis on House Sale Prices in King County. They will use the results of the study to advise their customers.

Questions:

1. What are the main predictors for House Sale Price?
2. Create a model to predict the House Sale Price.
3. Do house renovation affect the Sale Price?

Method

1. Data
    * Load
    * Explore
    * Clean
    * Check correlation
3. Model - Apply multiple regression and do validation
    * Baseline model with one best predictor
    * Second and Third models (Search for the next best predictors)
    * Final Model with five best predictors
3. Interpret
    * Interpret Final Model
    * Check multiple regression assumptions
4. Future work

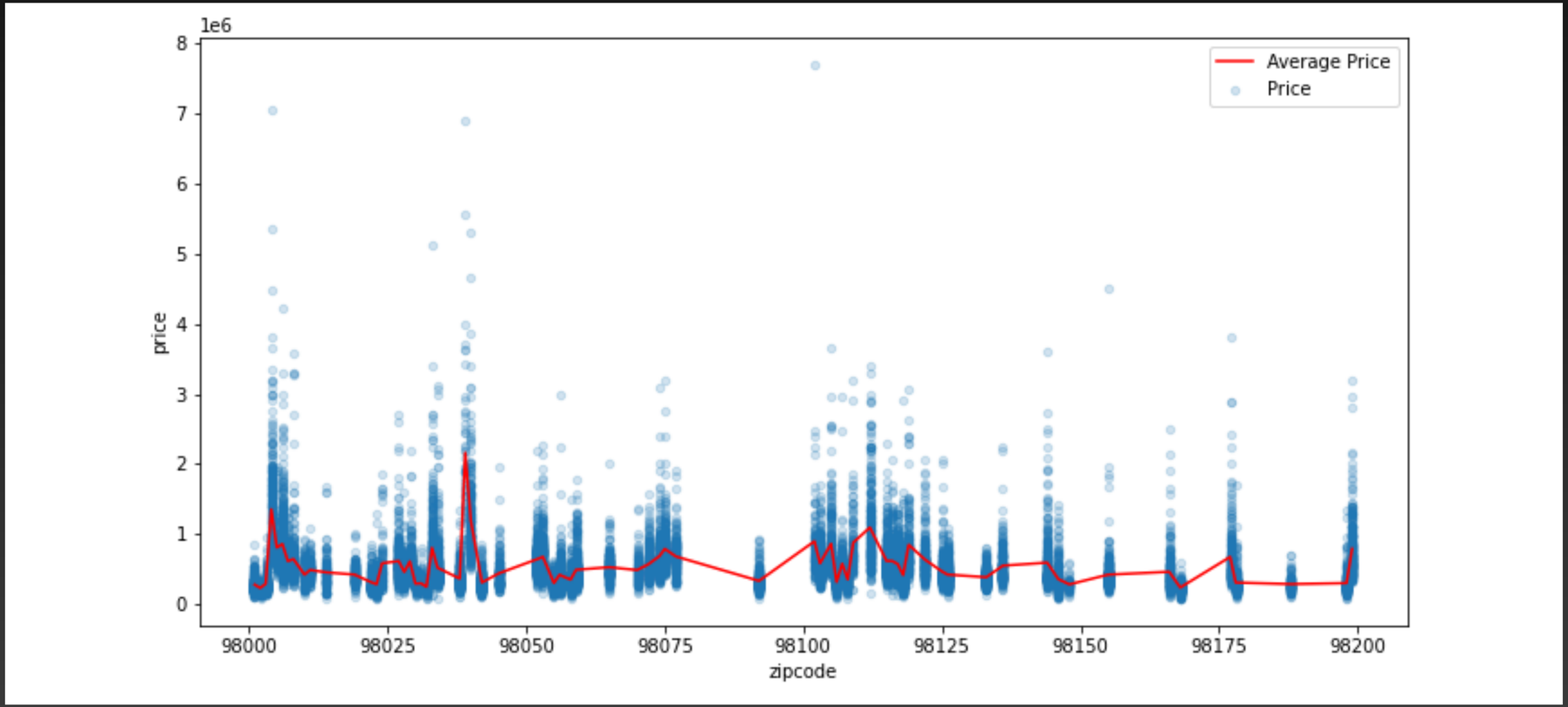


Data: Load & Explore & Clean

- King County House Sales Data ‘kc_house_data.csv'
- 21597 houses sold between May 2014 - May 2015
- 21 columns in unprocessed data
- 8 columns to be removed
- The columns left: 'id', 'price', 'bedrooms', 'bathrooms', ‘sqft_living', ‘sqft_lot’, 'floors', 'waterfront', 'condition', 'grade', ‘yr_built’, ‘yr_renovated', ‘zipcode'
- The missing values in ‘waterfront' and ‘yr_renovated’ are handled

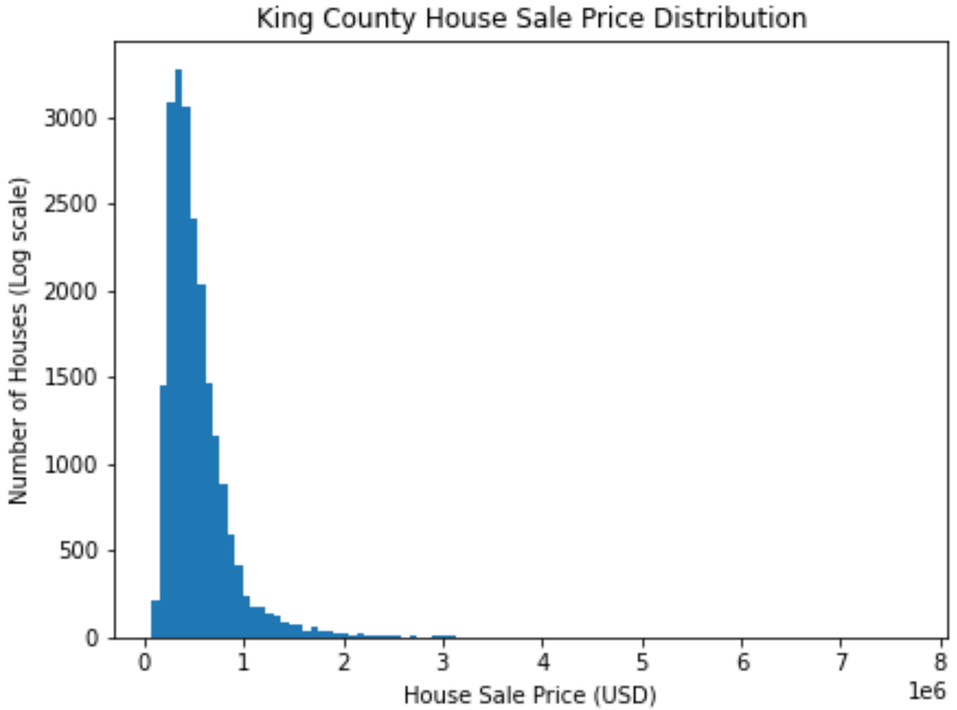
Data: ‘zipcode’

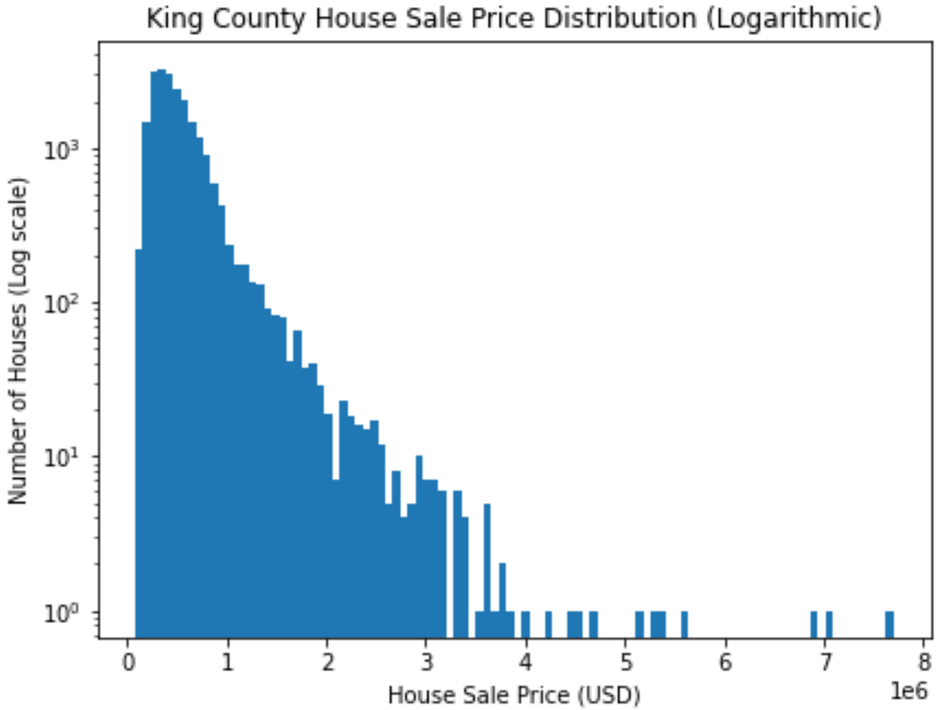
- The ‘zipcode’ is a categorical variable, and there are 69 zip codes in data
- The location is an important predictor in house sale price
- Hot-encoding is applied and dummy zipcodes columns are created



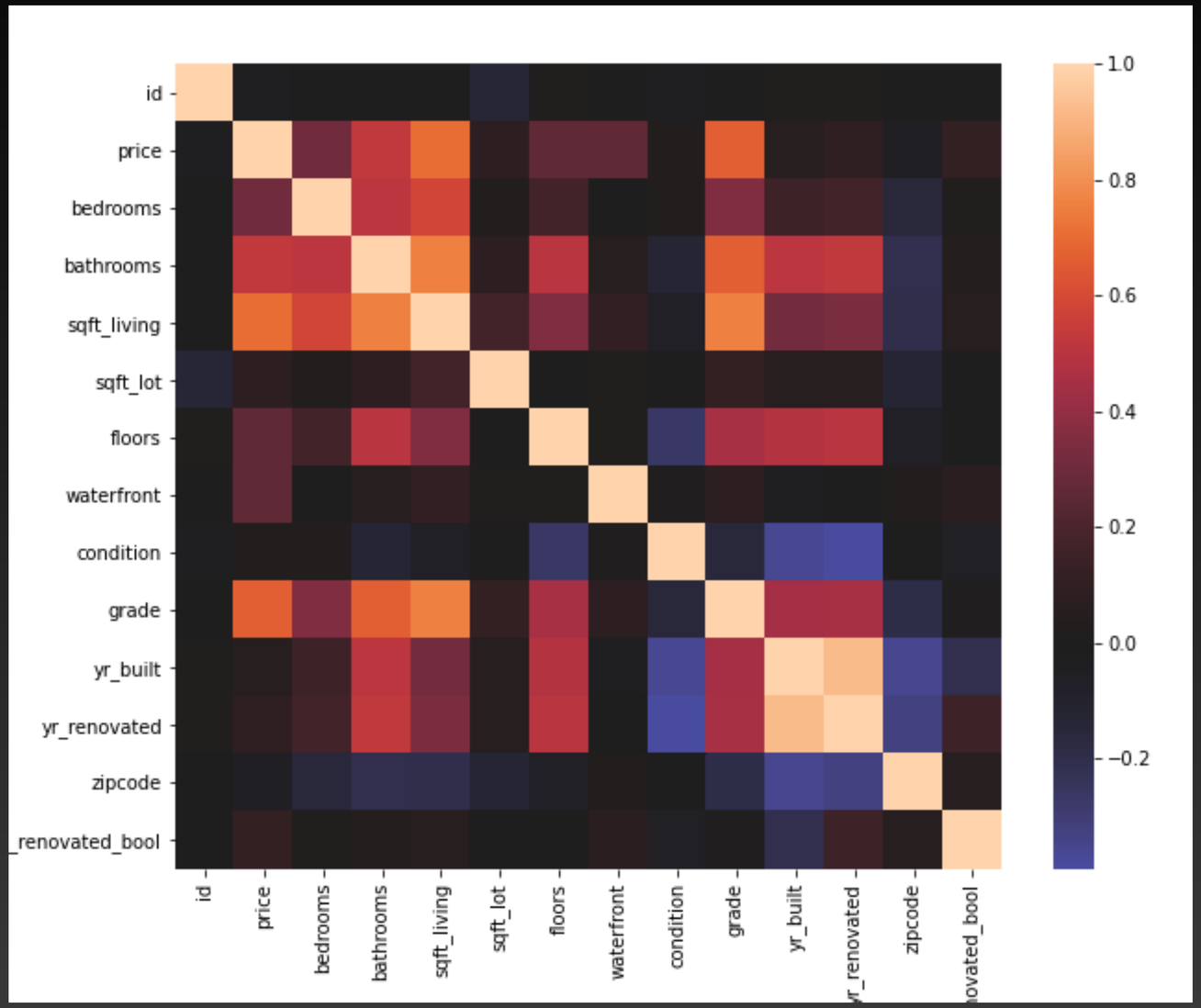
Data: ‘price’

- House sale ‘price’ is the target variable
- It is left-skewed normal distribution
- The outliers are observed in high sale prices





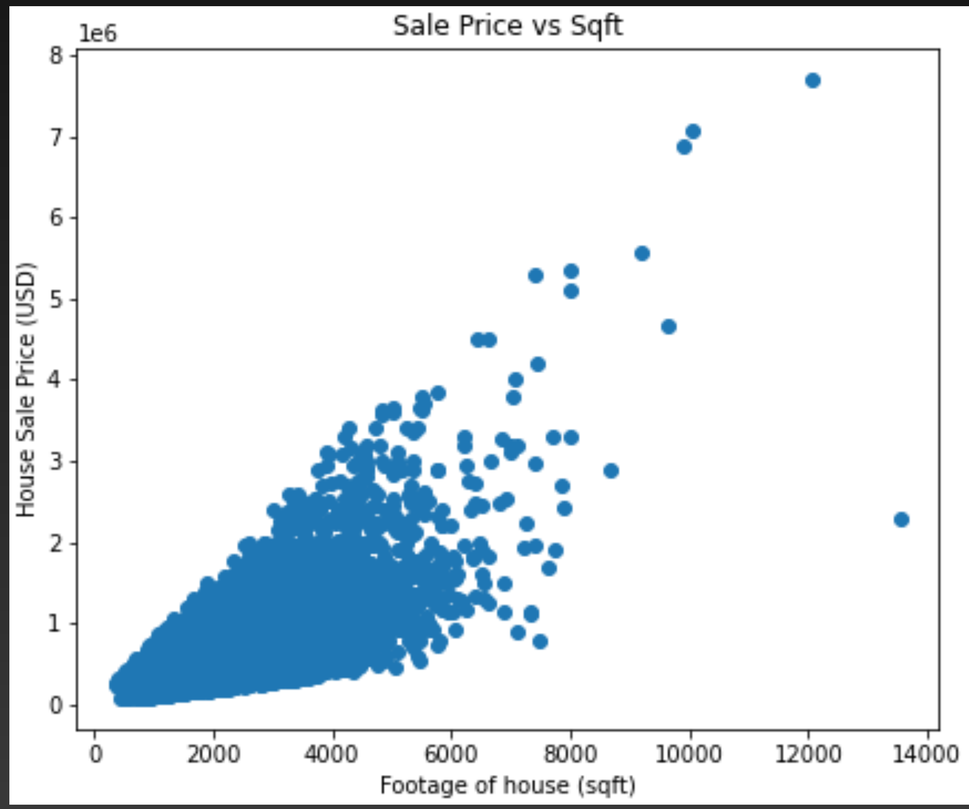
Data: Correlation

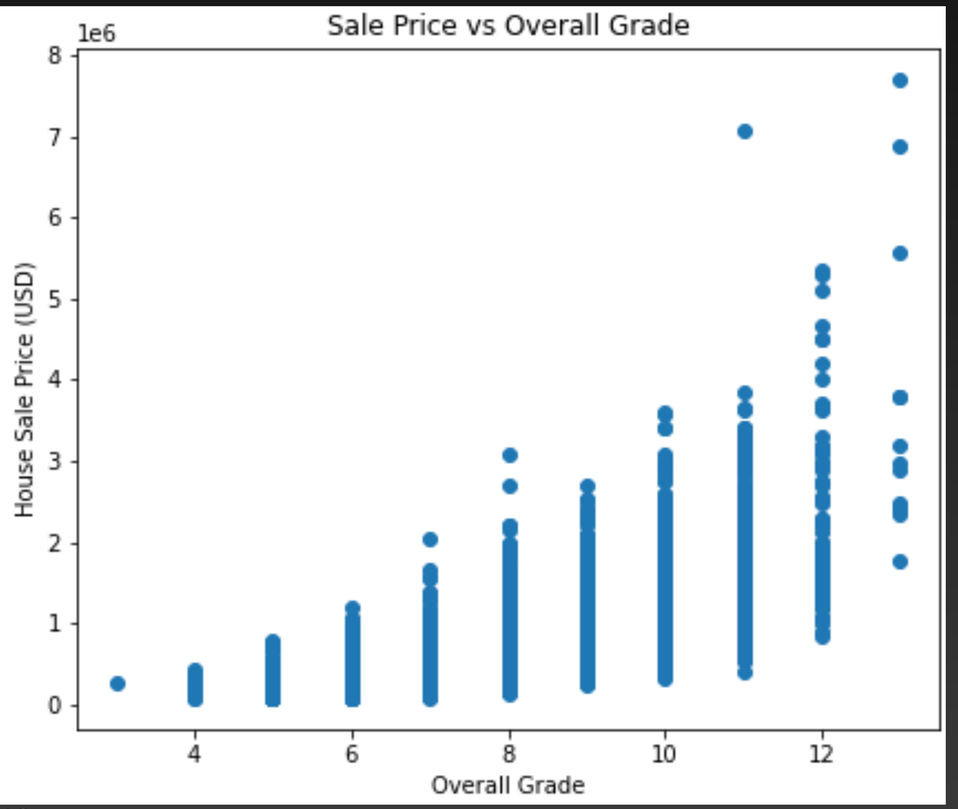


- good correlation between Price and (bedrooms, bathrooms, sqft_living, grade)
- good correlation among bedrooms, bathrooms, sqft_living, grade (multicollinearity)
- high correlation between yr_renovated and yr_built (multicollinearity)
- highest correlation with Sale Price: 'sqft_living' and 'grade'

Data: ‘sqft_living’ and ‘grade’

- sqft_living: square footage of the home (continuous variable)
- grade: overall grade given to the housing unit, based on King County grading system (categorical varible)

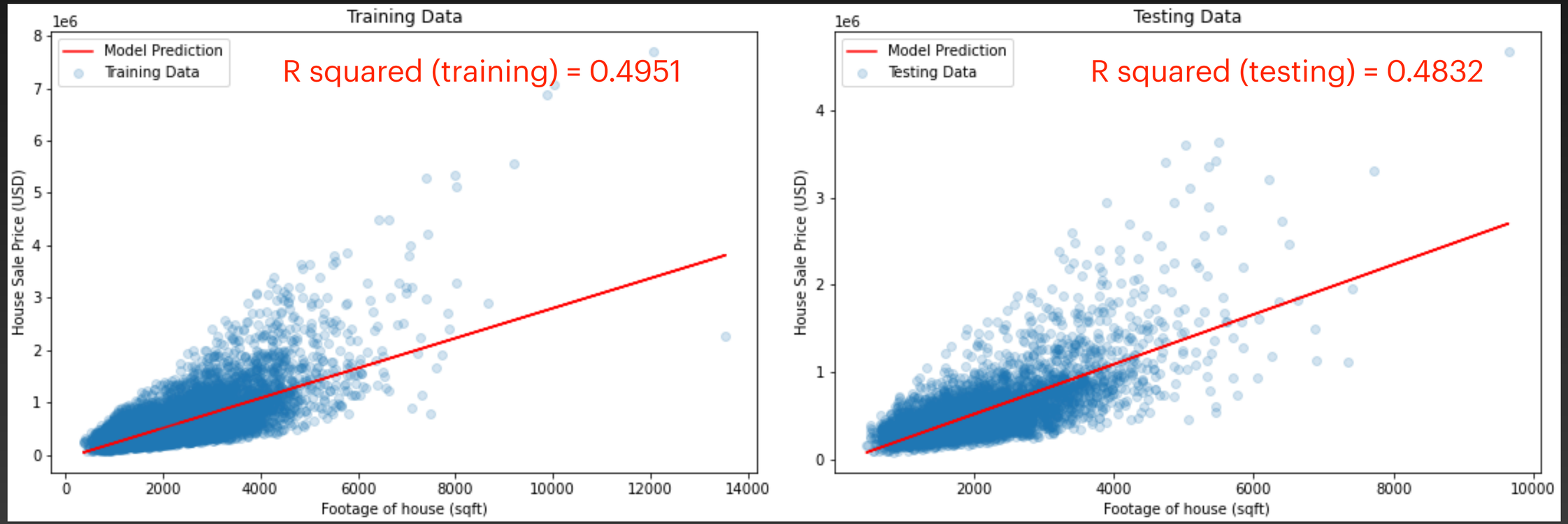




Model: baseline model

- Three steps in modeling:
  - Separate data into train and test splits
  - Apply Linear Regression Fit to training data and make predictions
  - Validate model on testing data

- Target: ‘price’ & Predictor: ‘sqft_living’
- Linear Regression fit results:
  - Slope: 285.58
  - y-intercept: -53321.49



Model: Second and Third models

Second model with two predictors:

	two_predictors	R_squared_train	R_squared_test
0	[sqft_living, zip_98004]	0.533440	0.520127
1	[sqft_living, waterfront]	0.529563	0.528285
2	[sqft_living, yr_built]	0.526124	0.518243
3	[sqft_living, zip_98039]	0.517281	0.506499
4	[sqft_living, yr_renovated]	0.516954	0.507128
5	[sqft_living, zip_98112]	0.511713	0.506115
6	[sqft_living, zip_98040]	0.508561	0.496405
7	[sqft_living, zip_98023]	0.504133	0.492635
8	[sqft_living, zip_98038]	0.502710	0.491157
9	[sqft_living, zip_98042]	0.502537	0.492249
10	[sqft_living, zip_98105]	0.502214	0.491189
11	[sqft_living, zip_98092]	0.501713	0.491271
12	[sqft_living, yr_renovated_bool]	0.501480	0.491107
13	[sqft_living, zip_98119]	0.501165	0.493318
14	[sqft_living, condition]	0.501093	0.489588

Third model with three predictors:

	three_predictors	R_squared_train	R_squared_test
0	[sqft_living, zip_98004, waterfront]	0.569135	0.566727
1	[sqft_living, yr_built, zip_98004]	0.562154	0.552604
2	[sqft_living, yr_built, waterfront]	0.556858	0.558546
3	[sqft_living, zip_98039, zip_98004]	0.556583	0.544304
4	[sqft_living, zip_98004, yr_renovated]	0.553657	0.543087
5	[sqft_living, zip_98039, waterfront]	0.551630	0.552382
6	[sqft_living, zip_98112, zip_98004]	0.551043	0.544249
7	[sqft_living, waterfront, yr_renovated]	0.549054	0.549292
8	[sqft_living, zip_98040, zip_98004]	0.548339	0.534251
9	[sqft_living, yr_built, zip_98039]	0.546934	0.539806
10	[sqft_living, zip_98112, waterfront]	0.546886	0.552160
11	[sqft_living, zip_98040, waterfront]	0.541900	0.539191
12	[sqft_living, zip_98023, zip_98004]	0.541846	0.528913
13	[sqft_living, zip_98105, zip_98004]	0.541012	0.528563
14	[sqft_living, zip_98038, zip_98004]	0.540323	0.527330

Best five predictors:

- sqft_living,
- waterfront,
- yr_built,
- zip_98004,
- zip_98039

Model: Final Model

- Multiple Linear Regression is applied on data
- Five Final model predictors:
  - 'sqft_living',
  - 'waterfront',
  - 'yr_built',
  - 'zip_98004',
  - 'zip_98039'
- R squared:
  - Training: 0.6159
  - Testing: 0.6179
- MSE = 4.549e+10 & RMSE = 2.133e+05

Results:

- Intercept: 4014990.7576159136
- Coefficients:
  - sqft_living: 2.844667e+02
  - waterfront: 8.040627e+05
  - yr_built: -2.071868e+03
  - zip_98004: 6.201505e+05
  - zip_98039: 1.162237e+06

Model: ‘yr_renovated’

Linear Regression Results:

- predictor: 'yr_renovated'
  - R squared for Training: 0.0102
  - R squared for Testing: 0.0061

- predictors: 'sqft_living', 'waterfront', 'yr_built', 'zip_98004', 'zip_98039', ‘yr_renovated' (final model predictors + ‘yr_renovated’)
  - R squared for Training: 0.6167
  - R squared for Testing: 0.6181

House Renovation doesn't have significant effect on House Sale Price.

Interpret: House Sale Price Prediction

y = b + m1x1 + m2x2 + m3x3 + m4x4 + m4x5

House Sale Price = 4014990.7576159136 + (2.844667e+02 * sqft_living) + (8.040627e+05 * waterfront) + (-2.071868e+03 * yr_built) + (6.201505e+05 * zip_98004) + (1.162237e+06 * zip_98039)

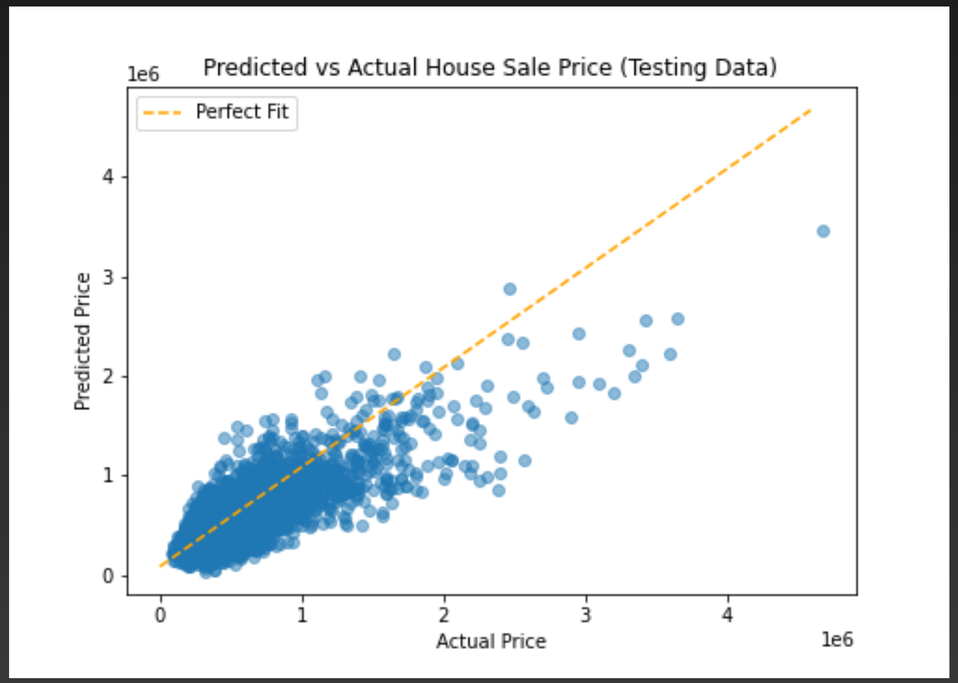
House Sale price prediction examples:

1. sqft_living=2800, waterfront=0, year_built=1978, zipcode_98004=0, zipcode_98039=0 => price = $713,342
2. sqft_living=4200, waterfront=0, year_built=1978, zipcode_98004=0, zipcode_98039=0 => price = $1,111,595
3. sqft_living=2800, waterfront=1, year_built=1978, zipcode_98004=0, zipcode_98039=0 => price = $1,517,405
4. sqft_living=2800, waterfront=0, year_built=2015, zipcode_98004=0, zipcode_98039=0 => price = $636,683
5. sqft_living=2800, waterfront=0, year_built=1978, zipcode_98004=1, zipcode_98039=0 => price = $1,333,493
6. sqft_living=2800, waterfront=0, year_built=1978, zipcode_98004=0, zipcode_98039=1 => price = $1,875,579

Interpret: Assumptions

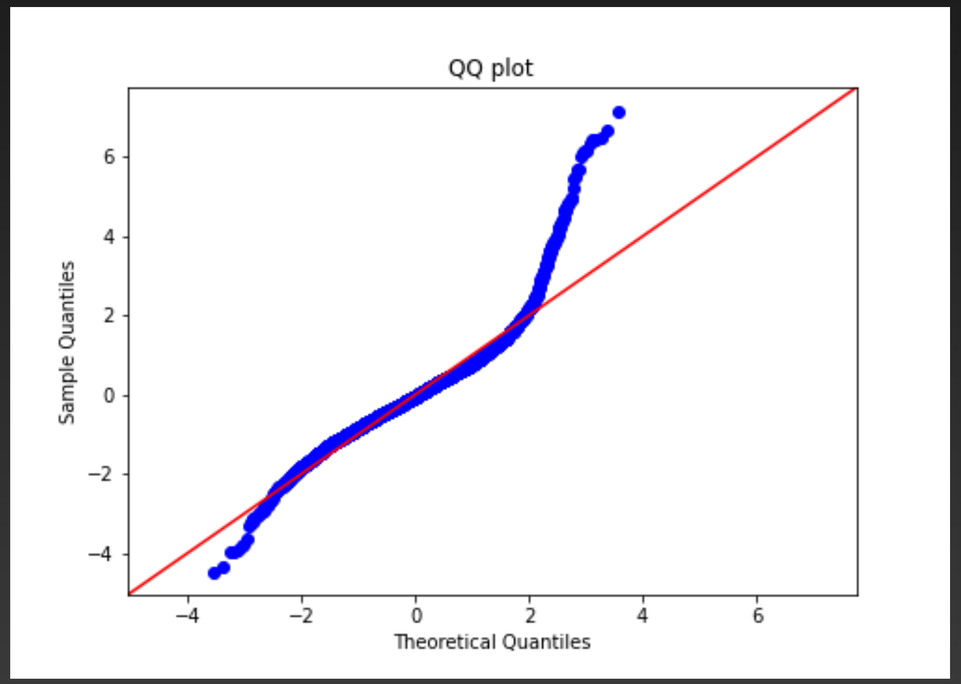
Linearity

- Linearity assumption holds for the majority of the data, except outliers at high sale prices
- At high sale prices, the predicted value is deflecting away from the perfect fit: outliers



Normality

- Normality assumption holds for the majority of the data, except outliers at high sale prices.
- skewness on tails/edges is caused by the outliners at the high sale prices.



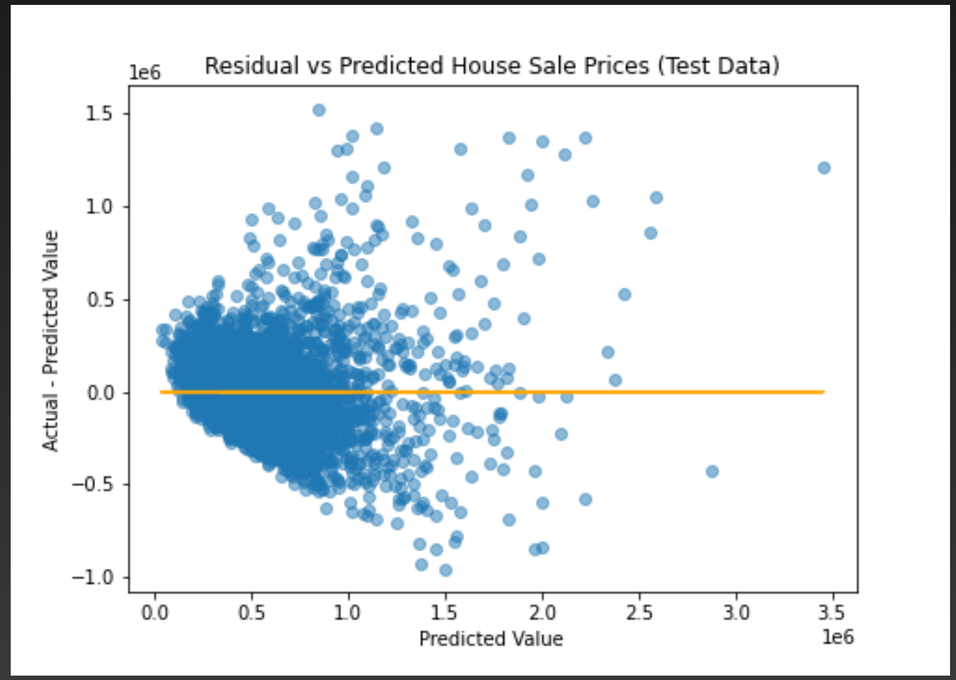
Interpret: Assumptions

Multicollinearity (Independence)

- Independence Assumption is violated since significant multicollinearity is observed
- Multicollinearity scores:
  - sqft_living: 6.460982
  - waterfront: 1.019214
  - yr_built: 6.320784
  - zip_98004: 1.029405
  - zip_98039: 1.010471
- sqft_living and yr_built are correlated!

Homoscedasticity

- Homoscedasticity assumption is violated
- The cone/funnel shape is observed on data.



Future Work

- Study outliers:
  - I guess one of the main causes for the assumption violations is outliers.
  - Remove from analysis?
- Study correlation between sqft_living and yr_built:
  - The multicollinearity caused by their correlation affects the model.
  - Should 'yr_built' be removed from model? Advantages and disadvantages?
- Explore Homoscedasticity:
  - Homoscedasticity is observed at low house sale prices as well as high.
  - How can we avoid it?

Questions?