House Sales in King County, USA

Yuka Chen
Coursera – Data Analysis with Python

This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.

Ridge regression model indicates that approximately 70.03% of the variance in the target variable can be explained by the model.

The dataset include 21 columns:

Variable	Description
id	A notation for a house
date	Date house was sold
price	Price is prediction target
bedrooms	Number of bedrooms
bathrooms	Number of bathrooms
sqft_living	Square footage of the home
sqft_lot	Square footage of the lot
floors	Total floors (levels) in house
waterfront	House which has a view to a waterfront
view	Has been viewed
condition	How good the condition is overall
grade	overall grade given to the housing unit, based on King County grading system
sqft_above	Square footage of house apart from basement
sqft_basement	Square footage of the basement
yr_built	Built Year
yr_renovated	Year when house was renovated
zipcode	Zip code
lat	Latitude coordinate
long	Longitude coordinate
sqft_living15	Living room area in 2015(implies-- some renovations) This might or might not have affected the lotsize area
sqft_lot15	LotSize area in 2015(implies-- some renovations)

Each column’s type:

```
Unnamed: 0         int64
id                 int64
date              object
price            float64
bedrooms         float64
bathrooms        float64
sqft_living        int64
sqft_lot           int64
floors           float64
waterfront         int64
view               int64
condition          int64
grade              int64
sqft_above         int64
sqft_basement      int64
yr_built           int64
yr_renovated       int64
zipcode            int64
lat              float64
long             float64
sqft_living15      int64
sqft_lot15         int64
dtype: object
```

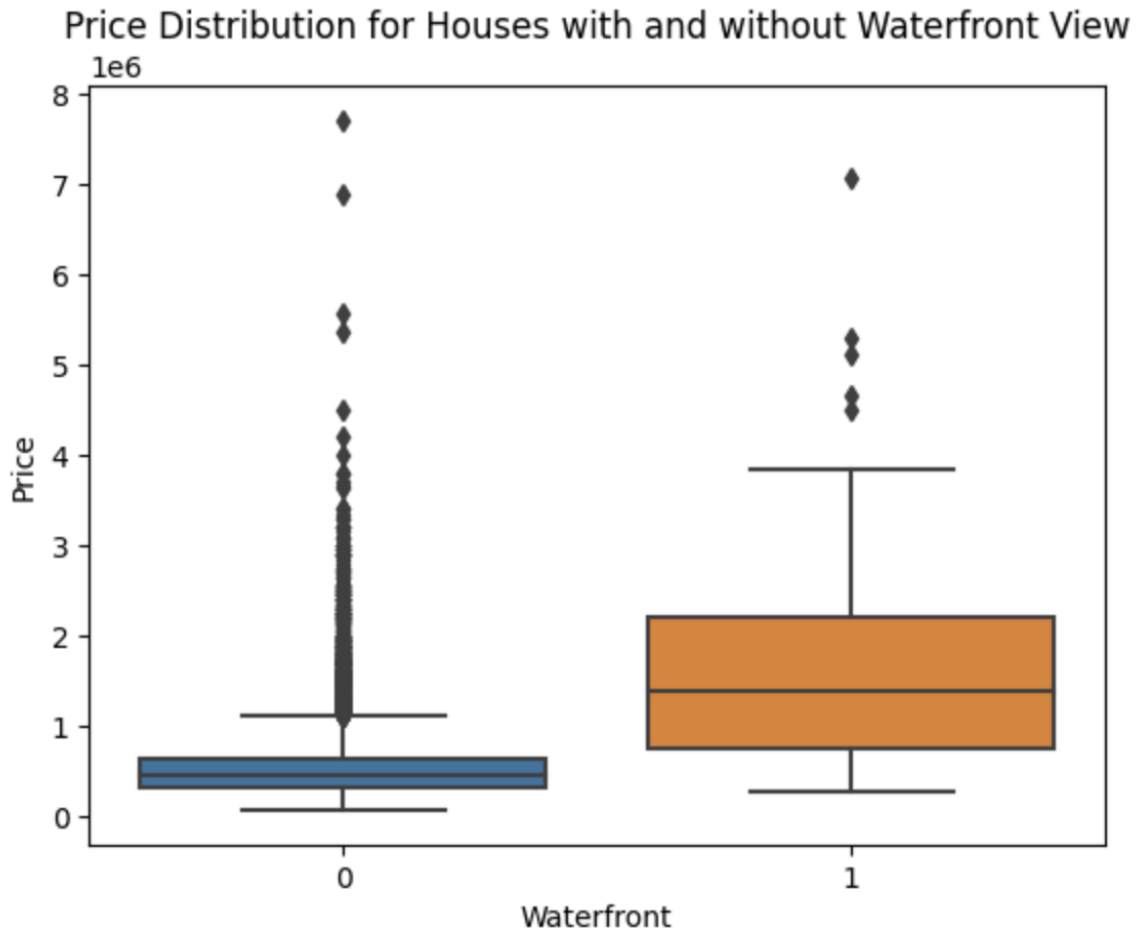
the ‘id’ and the ‘unnamed: 0’ will be removed for the model.

Since there are some missing value in bedrooms and bathrooms, we would replace the missing value with the average.

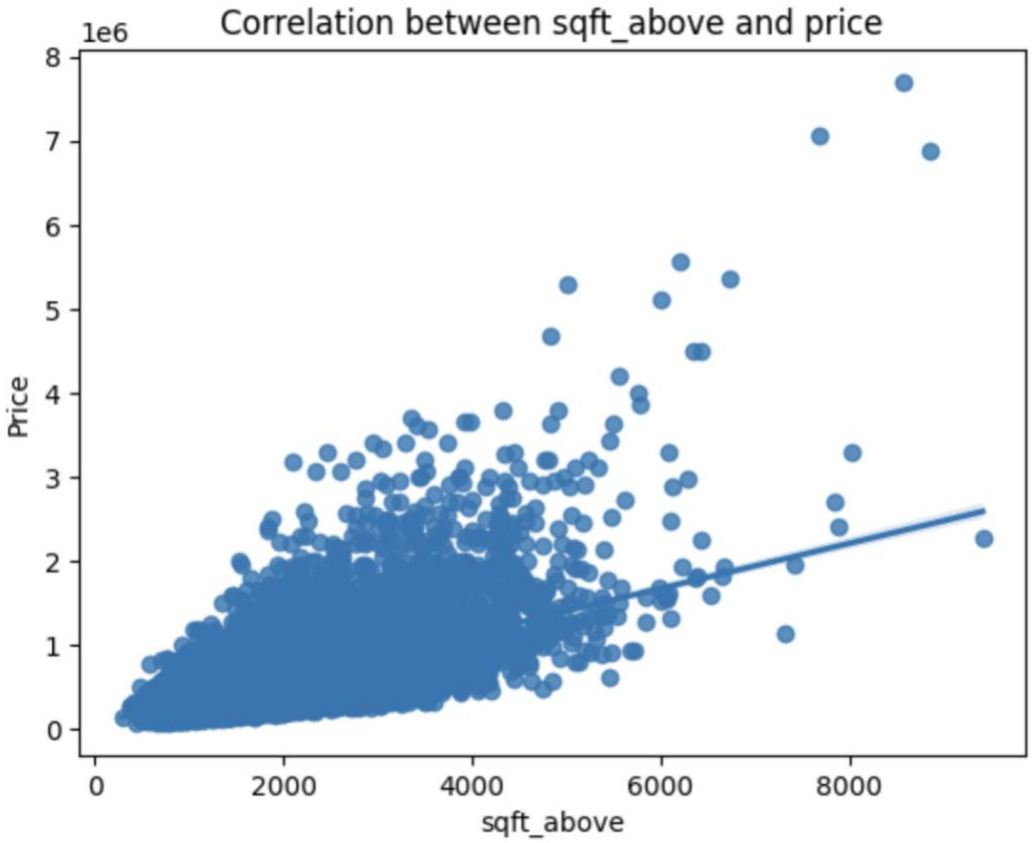
Exploratory Data Analysis

Majority of the houses in King County has one floors. The second one houses have two floors. And third majority of houses have 1.5 floors.

	floors
1.0	10680
2.0	8241
1.5	1910
3.0	613
2.5	161
3.5	8



From the plot we could see that the houses with waterfront has higher sales prices than the ones do not have waterfront.



The larger square footage of house apart from basement (sqft_above), the more expensive the house is.

```
zipcode         -0.053203
long             0.021626
condition        0.036362
yr_built         0.054012
sqft_lot15       0.082447
sqft_lot         0.089661
yr_renovated     0.126434
floors           0.256794
waterfront       0.266369
lat              0.307003
bedrooms         0.308797
sqft_basement    0.323816
view             0.397293
bathrooms        0.525738
sqft_living15    0.585379
sqft_above       0.605567
grade            0.667434
sqft_living      0.702035
price            1.000000
Name: price, dtype: float64
```

We can see that the number of bathrooms, sqft_living, sqft_above, and grade are corrated with the house price.

Model Development:

1) **Simple linear regression with bathrooms**

First, we build a simple model with one features.

$$\widehat{price} = 8930.37 + 251051 * bathrooms$$

In the given model, the coefficient for the bathroom numbers is 251051. This means that for every one-unit increase in the bathroom numbers, the predicted price is expected to increase by $251051.

Conversely, for every one-unit decrease in the bathroom numbers, the predicted price is expected to decrease by $251051.

This model gives us 0.276 R-square. An R-squared value of 0.27639 for simple regression model indicates that approximately 27.6% of the variance in the target variable can be explained by the model.

2) **Simple linear regression with square footage of the home (sqft_living)**

$$\widehat{price} = -43580 + 280.6 * sqft\_living$$

In the given model, the coefficient for the sqft_living variable is 280.6. This means that for every one-unit increase in the sqft_living variable (square footage of the home), the predicted price is expected to increase by $280.6.

Conversely, for every one-unit decrease in the sqft_living variable, the predicted price is expected to decrease by $280.6.

This model gives us an R-square = 0.493. An R-squared value of 0.493 for simple regression model indicates that approximately 49.3% of the variance in the target variable can be explained by the model.

3) **With multiple features**

1. Floors
2. Waterfront
3. Lat
4. Bedrooms
5. sqft_basement
6. view
7. bathrooms
8. sqft_living15
9. sqft_above
10. grade
11. sqft_living

$$\begin{aligned}\widehat{price} = {} & -32329161.95 - 26658.22 * \text{Floors} + 615651.69 * \text{waterfront} + 671773.09 \\ & * \text{lat} - 27756.655 * \text{bedrooms} - 2689224340000000 * \text{sqft\_basement} \\ & + 67022.54 * \text{view} - 4440.514 * \text{bathrooms} + 2.162 * \text{sqft\_living15} \\ & - -2689224340000000 * \text{sqft\_above} + 82029.24 * \text{grade} \\ & + 2689224340000000 * \text{sqft\_living}\end{aligned}$$

The R-square for this model is: 0.658, which indicates that approximately 49.3% of the variance in the target variable can be explained by the model.

Model Evaluation and Refinement

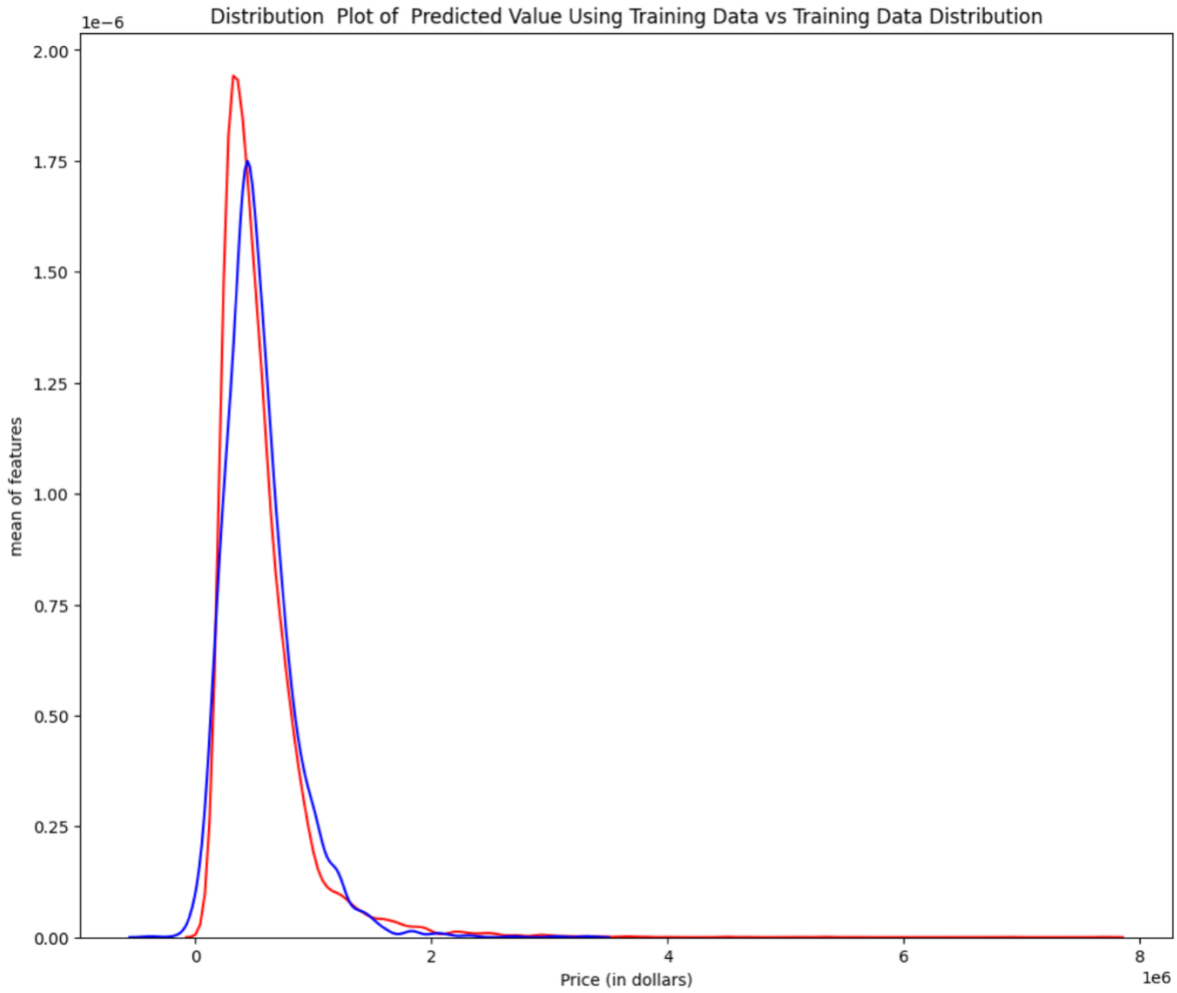
Split the model into training and testing sets.

- number of test samples: 3242

- number of training samples: 18371

Create and fit a Ridge regression object using the training data, set the regularization parameter to 0.1.

We got 0.648 R-square, which indicates that approximately 64.8% of the variance in the target variable can be explained by the model. It also represents the accuracy of your multiple features Ridge regression model



Polynomial transformation is a process of creating polynomial features from the original input features in a model. It involves generating new features by taking powers and combinations of the original features.

In a polynomial transformation, each input feature is raised to various powers (e.g., squared, cubed, etc.) and combined to create new polynomial features. This allows the model to capture nonlinear relationships between the features and the target variable, enabling it to fit more complex patterns in the data.

The polynomial transformation can be applied to different degrees. For example, a second-degree polynomial transformation for a single input feature x would create new features x^2, x^3, etc. If there are multiple input features, combinations of the features are also generated. For instance, with two input features x and y, the second-degree polynomial transformation would include features like x^2, y^2, xy, etc.

By introducing these polynomial features, the model becomes more flexible and capable of capturing nonlinear relationships. It can help improve the model's accuracy and enable it to fit curved or nonlinear patterns in the data. However, it's important to note that applying polynomial transformation can also increase the complexity of the model and the risk of overfitting if not carefully controlled.

With the second degree of polynomial transformation, model’s r-square increased to 0.70, which has better accuracy rate.