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 coloumns:

A picture containing text, receipt, screenshot, number

Description automatically generated

Each column’s type:

A screenshot of a computer code

Description automatically generated with low confidence

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.

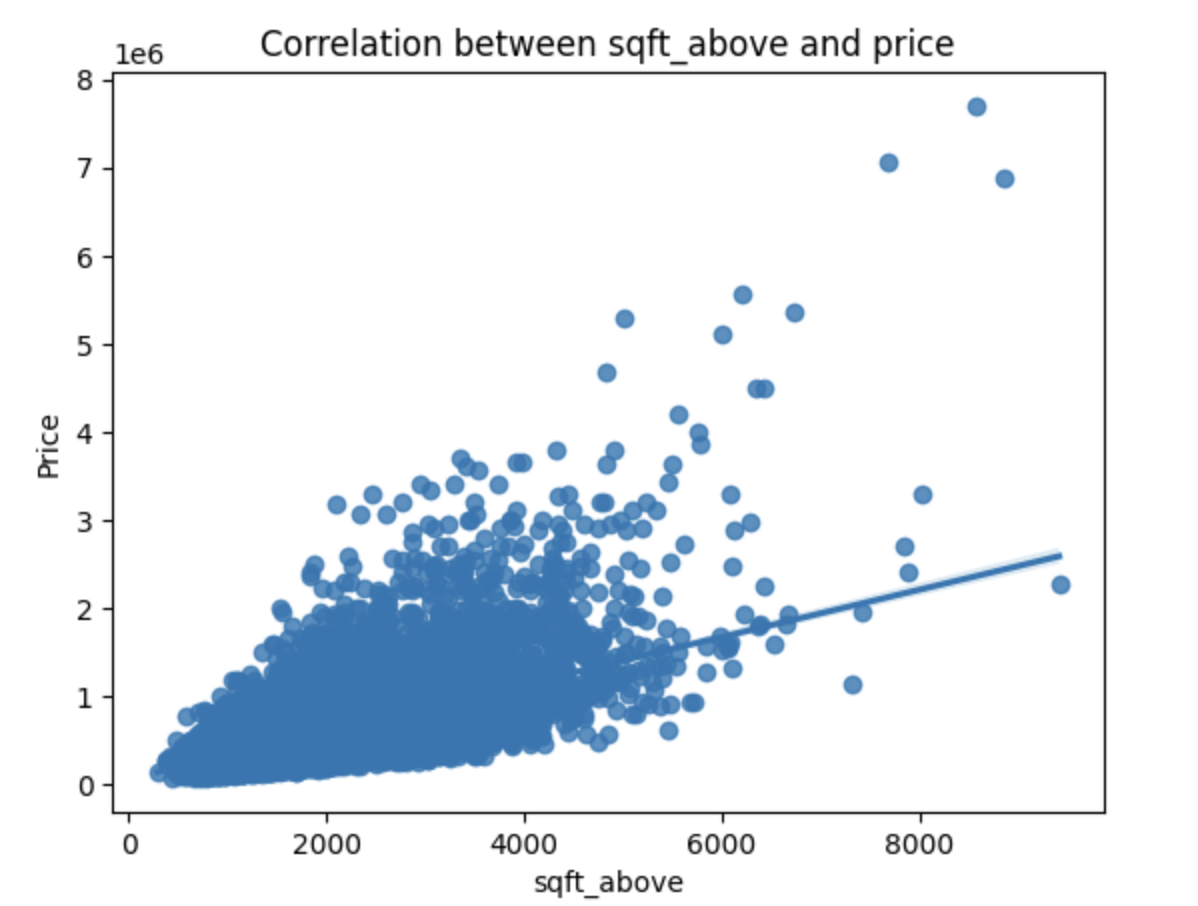
A screenshot of a cell phone

Description automatically generated with low confidence

A picture containing text, diagram, screenshot, line

Description automatically generated

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.

A screenshot of a computer

Description automatically generated with low confidence



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.

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.

1. **Simple linear regression with square footage of the home (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.

1. **With multiple features**
2. Floors
3. Waterfront
4. Lat
5. Bedrooms
6. sqft\_basement
7. view
8. bathrooms
9. sqft\_living15
10. sqft\_above
11. grade
12. sqft\_living

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

A picture containing text, screenshot, plot, diagram

Description automatically generated

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.