Dataset:
The dataset has been collected from Kaggle.com . The data was scraped from publicly available results posted every week from Domain.com.au. The dataset includes Address, Type of Real estate, Suburb, Method of Selling, Rooms, Price, Real Estate Agent, Date of Sale, and distance from CBD. There are total of 13580 observations and 21 columns.

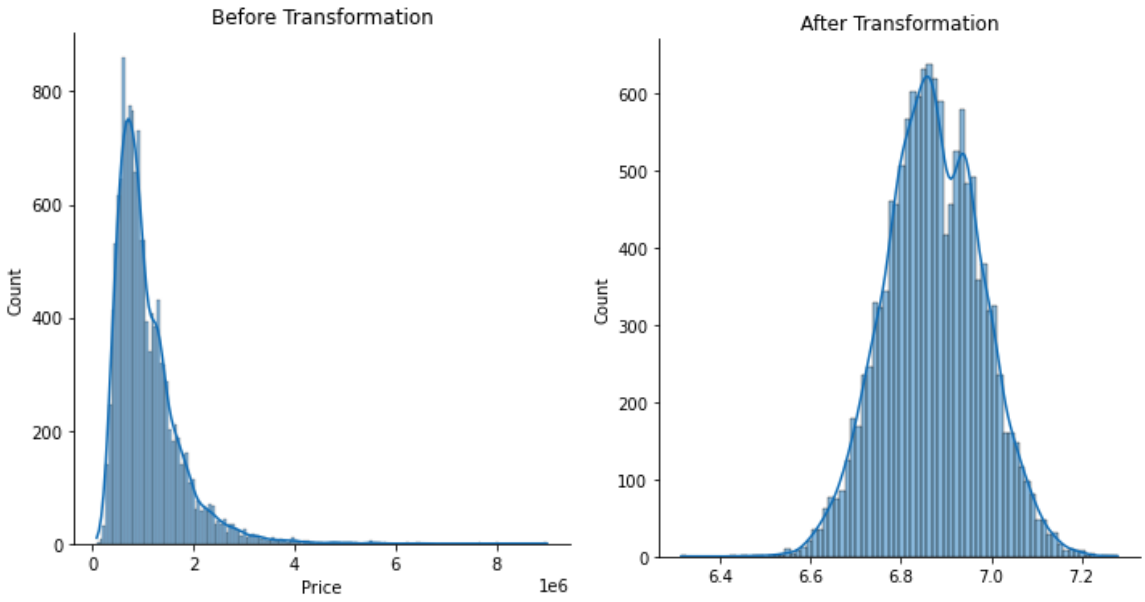
Goal:
The goal of this analysis would be to come up with a regression model which takes in input various features (20 of them) and try to accurately predict the price of the house.

Pre-processing steps:
The first step was to get some valuable information of the dataset using describe() and info() methods of the pandas library. After which I have calculated the number of missing values each feature have. It was found that BuildingArea and the YearBuilt feature had nearly more than 40% missing values. Hence, I deleted those features completely. I also deleted the Address feature as this would covered up by the latitude and longitude variables. Then I filled up very few missing values in various columns by mean for numerical feature and by most common value for the categorical variable. I also converted the Date variable by keeping only the year value. Next step into the pre-processing step was converting the categorical data into one-hot representation.

Linear Regression Model:

1. **Straightforward Linear Regression Model:** in this step, I created a Linear Regression model without any advance techniques being used. The dataset was divided into X and y from which both of the arrays were split into train and test set. After performing Linear Regression, the R2 score was 0.5686 which served as baseline for other method being used.

2. **Standardised Variables:** In the second method, the predictor variables were standardised and the y variable was normalised using box-cox transformation. After performing the transformation, I fit a linear model for the data and the step did some improvement in the performance. The R2 score for this step was 0.6079



3. **Ridge Regression Model:** In the third method, I applied the ridge regression model using alpha values [0.005, 0.05, 0.1, 0.3, 1, 3, 5, 10, 15, 30, 80] and cv=4. The model did not do better than the Standardised variables and R2 score for this model is 0.5563

The best model out of the three performed was Standardising the predictor variables and normalising the output variable and fitting a Multiple Linear Regression Model.

Summary

On the basis of the observations present into the dataset, it is possible to predict the price of the house with a R2_score of 0.60 and also from the scatter plot of our prediction and the test dataset, we can conclude that the model did a good job fitting the data considering it to be a linear regression model.

Possible Flaws:

The dataset could have added features describing the population density of the locality, the distance from public transport, the availability of various entertainment options, shopping malls, crime rate. These features highly influence the price of a property in a locality. If those features were present then, most probably the model could have performed much better in predicting the price of the house. This are the flaws I would like to solve if I get a chance in the future with some information availability.