TASK1

Introduction

A house price has a significant impact on the choice to buy it. In order to better understand the King County, USA housing market and estimate a price for a house this analysis will employ simple linear regression to examine the link between the price and other variables contained in the dataset. I advanced from using a single feature to multiple features. The Task1.py file attached contains the code.

Keynotes

There are 18 features (independent variable) and 1 target variable (dependent variable).

• The target/dependent variable is the price (y)
• The features are bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, grade, sqft_above, sqft_basement, yr_built , yr_renovated , zipcode ,lat,long,sqft_living15,sqft_lot15.
• Intercept: When the independent variable (X) is equal to zero, this is the predicted mean value for the dependent variable (y).
• Coefficient: The magnitude of each independent variable's coefficient indicates the extent of its influence on the dependent variable. In simple linear regression, the coefficient indicates whether the dependent variable is predicted to go up or down by one, depending on whether it is positive or negative.
• Mean Squared Error: This demonstrates how near a set of points a regression line is.
• Coefficient of determination: The percentage of variance in the dependent variable that can be explained or predicted by the independent variable is known as the R squared or the coefficient of determination. The square of the correlation between the dependent and independent variables can also be used to describe it.
• Simple linear regression: Is a regression model that uses a straight line to calculate the connection between one independent variable and one dependent variable.

STEP1

The first step is to Import the essential Python functions and libraries to help with the development of this algorithm. Read the dataset from a CSV file, this provides information on the dataset's rows and columns. This dataset contains 21613 entries and 19 columns.

The correlation of features is represented using a heatmap. The light colour in the heatmap indicates a strong positive correlation, and the dark colour indicates a negative correlation (can be seen in task1.pyfile attached). The sqft_living has the highest positive correlation of 0.7 to price. With grade, sqft_above, sqft_living15, and bathroom also having a positive correlation. I removed other features using the highest positive correlation to plot our simple linear regression.

STEP 2

In this phase, the dataset is divided into one-third (1/3) of the original dataset for testing and 2/3 for training. We then conduct the computations and plot the data to observe a

much more precise conclusion. In order to obtain the regression line, the linear regression function would also be performed on the training data. The visualization is below:

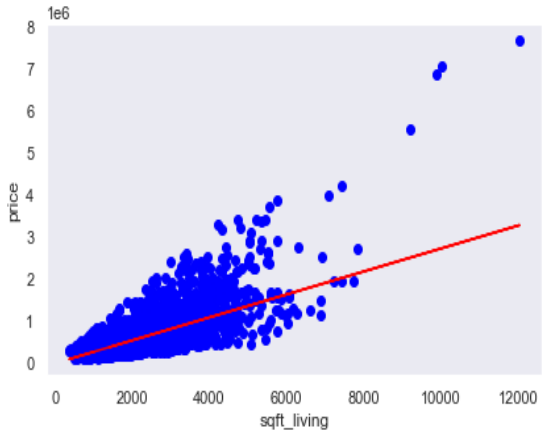

Fig1.

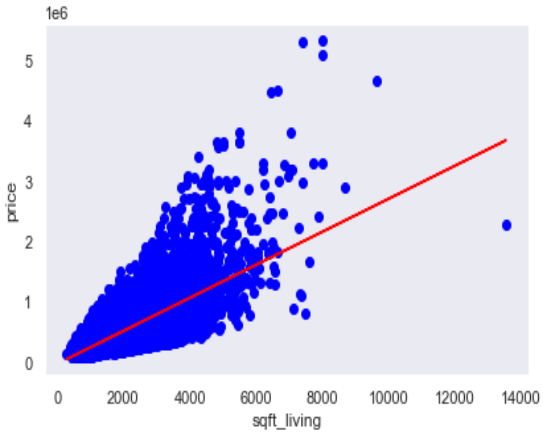

Fig2.

Including more features

Four features were added, however, their effects on the model's performance were quite minimal. The mean square error shrank, which marginally boosted the coefficient of determination. When all the features were taken into account, the model showed an improvement, as seen in the study below:(details in the task1.py)

	One features	Four features	All features
Mean squared error	72251932678.751 92	65890394607.029 335	40475314947.228 989
Coefficient of determination	0.500052	0.54	0.71

Effectiveness of the model

After training the data using 75% of the dataset and testing with 25% this helped to get our coefficient of determination(r2_score), using a single feature (price to sqft_living) the model was valued to be 50%. After using all features it was 71%. It was observed that as the features increased the error in the model reduced as shown in the table above (mean squared error).

How could you make further improvements?

Examining all the features in relation to house price gave a better coefficient of determination (71%) showing that the price of houses in King County, USA are better influenced considering all features. This also shows that if more features are added to this dataset the model might improve

Conclusion of the model

After correlation, it shows that price is best correlated with sqft_living as well as grade,sqft_above,sqft_living15, and bathroom. When making a decision it is best to consider all features in determining the price of a hose at King County, USA. The minimum and maximum prices of the house, according to the description, were £7500000 and £7700000, respectively. 33 bedrooms and 8 baths were the maximum number in the house that was sold. According to the 50% and 75% percentiles, the majority of sold homes had three or four bedrooms.