Predicting house prices using multiple linear regression

King County housing data set

Dataset

- The king county housing data set contains information about houses including; size, location, condition and other features.

Analysis Question

- The analysis aims towards identifying the most appropriate variables from the dataset that can create a multiple linear regression.
- The model will help home owners interested in selling their homes by informing them on the most important factors to consider in order to improve sale prices.

Data preparation

- The data was prepared for analysis by dealing with missing values and getting rid of outliers.
- One hot encoding was also performed on one of the columns which contained categorical data.

Correlation

- The features that were highly correlated with price were considered for inclusion in the model.
- The variables that were highly correlated with each other were excluded from the model.

Simple Linear Regression model

- Simple linear regression models were constructed with the selected variables to confirm that the assumptions of linear regression have been met.
- The variables were checked for linearity, normality and homoscedasticity.

Final Multiple Linear Regression model

$Y = 0.7079 + 0.0221 sqft\_living + 0.0102 sqft\_living15 - 0.0076 bedroom + \varepsilon$

- $R^2 = 0.488$
- The adjusted R squared value shows that the model accounts for 48.8% variability in price

Hypothesis Testing

- Null hypothesis: no relationship exists between the chosen explanatory variables and response variable.
- Alternative hypothesis: a relationship exists between the chosen explanatory variables and response variable.
- The model coefficients all display a p value 0.00 that is below 0.05
- The null hypothesis is rejected. This evidently shows that a relationship exists between the explanatory variables and response variable. The model is therefore statistically significant.

Multiple linear regression model

- Since the model has log transformed predictors and a log transformed target, the coefficient values are interpreted as percentages.
- A lack of change in sqft_living, bedrooms and sqft_living15, sale price would be 70.79%.
- A 1% increase in square footage of living space in the home is associated with a 2% increase in sale price.
- A 1% increase in number of bedrooms results in a 0.76% decrease in sale price.
- A1% increase in square footage of interior living space for the nearest 15 neighbors is associated with a 1% increase in sale price.

Conclusions

- Square footage of the living space, number of bedrooms and square footage of interior living space for the nearest 15 neighbours strongly predict house prices in a multiple linear regression.

Recommendations

- Expand square footage of living space
- Encourage the neighbours to expand square footage of living space and enhance interior design.
- Reduce the number of bedrooms.

Recommendations for further study

- Checking on the best predictors for price in other counties.
- Determining how analysis changes if extreme values are excluded in the model.