Lab 4: House price data

The variables in the Excel data set **Lab 4.xlsx** are as follows. The data[1] relates the selling price (US dollars) to characteristics of a house.

Variable Name	Description
price	Selling price
bedrooms	Number of bedrooms
bathrooms	Number of bathrooms/bedroom
sqft_living	Square footage of building
sqft_lot	Total square footage of lot being sold
floors	Number of floors
condition	Condition of the building (1 = Poor, 2 = Good, 3 = Excellent)
grade	Grade given to the property based on city grading system
yr_built	Year the property was built
lat	Latitude of the property
long	Longitude of the property

Question:

(i) Using appropriate plots explore associations between potential predictor variables selling price.

(ii) Fit a regression model to predict price by considering each of the following:

1. Use VIF statistics to identify any problems with multicollinearity

2. Transformation of the dependent variable price

3. Build a regression model that contains only significant predictor variables using a backward regression approach

4. Use a partial F-test to identify if the variable "condition" is a significant predictor in the regression model, controlling for all other variables in the model.

5. Use residual plots to check the whether the final model satisfies the assumptions of the regression.

6. Use Cook's D to identify if there are influential points.

7. What is the R^2 and Adjusted R^2 fit statistics for your final model?

(iii) Use your fitted regression model to determine the properties that got an usually high prices and an unusually low price.

[1] The data used is adapted from the Kaggle.com data set "House Sales in King's County, USA"