Below are some observations on the independent variables in the Leslie Salt data:

For the continuous variables, we have tried some logarithmic transformation, and considered it appropriate wherever it improved the linear correlation.

* Country: It’s a categorical variable with two labels – giving it a binary nature. Value of 1 means the location is Santa Clara and 0 means San Mateo. Around 61% data points are from Santa Clara. If we look at the summary – data in San Mateo shows higher median and wider range of prices.
* Size: The size of the property is slightly negatively related to the price – which is surprising. It seems that the other features have more importance than the size for this type of properties. The log transformation does increase the correlation slightly, however, we may go without any transformation given the negative relation is not practical and this variable seems to have very less importance.
* Elevation: This is a continuous variable with positive correlation of 0.35 with price. Log transformation does improve the correlation to 0.43, so we may consider this transformation.
  + Assumption: We considered zero for the cells with LN(0), which were showing #num error.
  + LN of price further improves the correlation (around 0.57)
* Sewer: This is a continuous variable with negative correlation of 0.39. If we take the LN of the price, the correlation improves to (-0.47).
* Date: This continuous variable (time elapsed since deal) is showing strong negative correlation with the price. For ease of modelling and interpretation we can take the absolute values. The correlation marginally improves when we take LN of price.
  + We tried converting the months in years by dividing by 12 - however, that did not impact the correlations.
* Flood: It’s a binary categorical variable. The summary shows much lower prices for properties subject to flooding. So it would be a good predictor.
* Distance: This is a continuous variable. The distance has a positive relation with the price – which is counterintuitive. While the relation is insignificant without transformation (correlation 0.09), it improves to 0.30 when the LN of distance is considered.

Based on the above observation, the regression will consider the following:

* Dependent variable = LN(price)
* Independent variable= Country, Size, LN(Elevation), Sewer, (abs) Date, Flood, LN(Distance)

When **no transformation** is used – we got a regression model with adjusted R squared of 0.67, and the model implies that Size, Date and Distance are not significant predictors according to the respective p values.

However, **with transformations** as mentioned above, the adjusted R squared improves to 0.80. The resulting model is shown below:

Ln (Price) = 2.726 – 0.198 \* Country – 7(E-05) \* Size + 0.3206 \* LN (Elevation) – 5(E-05) Sewer – 0.013 \* absolute Date – 0.849 \* Flood + 0.2176 \* Ln (Distance)



Given Country, Size and Sewer are insignificant predictors; we ran another model excluding these variables. The Adjusted R squared was almost unchanged. The model summary is shown below:

