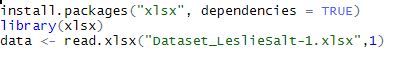
**Problem 2 – Leslie Salt Data Set:**

**Problem statement**: We have been given data on 31 byland properties near parcel of land owned by the Leslie Sal Company, that were sold during the previous 10 years. In addition to the transaction price for each property, data on large number of other factors, including size, time of sale, elevation, location, and access to sewers is also given. A description of the variables is provided below.

**Description of the variables**  
Price –> Sales price in $000 per acre  
County –> San Mateo=0, Santa Clara =1  
Size –> Size of the property in acres  
Elevation –> Average Elevation in foot above sea level  
Sewer –> Distance (in feet) to nearest sewer connection  
Date –> Date of sale counting backward from current time (in months)  
Flood –> Subject to flooding by tidal action =1; otherwise =0  
Distance –> Distance in miles from Leslie Property (in almost all cases, this is toward San Francisco

**1. Let us load and read the data:**



### 2. Transformation:

Here we are converting county and flood into factor variables with corresponding level

values being replaced by labels.

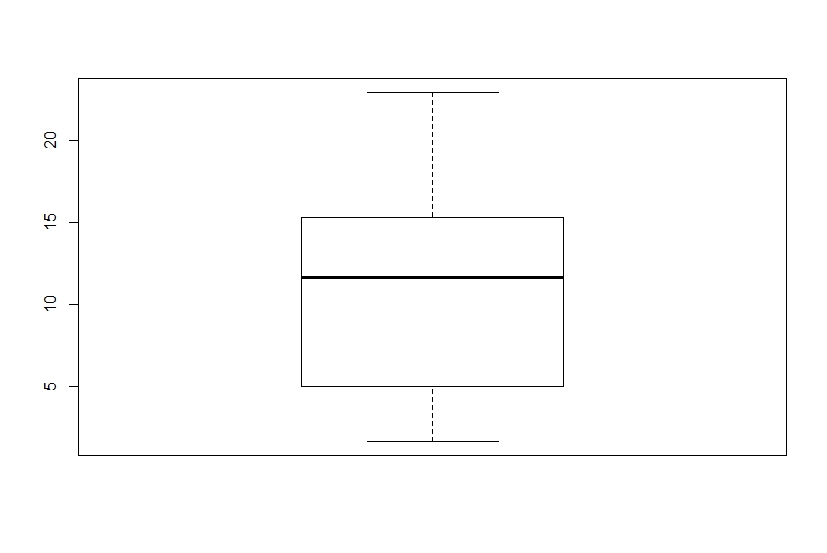


### 3. Let us do the data exploration through box plot to identify the outliers and correlations

### 

From the boxplot of **Price** we have an outlier which will impact the prediction. There is an outlier (a value of 37.20). This outlier could impact our model. Even if we remove this, we have sufficient sample data





4. **Assumptions:**

To predict a model with the variables given, we can do regression analysis.Before doing the regression we need to check if our dataset satisfies all the regression assumptions.

### 4.1.Multi collinearity:VIF - Variation inflation factor

### 

### 

### VIF is less than 10 hence there is no multi collinearity

### 4.2 Durbin Watson Test to test Auto Correlation

### 

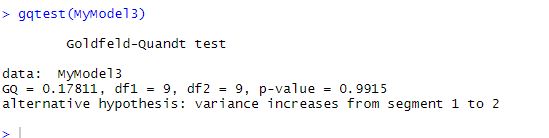
DW value 2 indicates there is no Auto correlation. Thumb rule 1.5 to 2.5 safe to conclude no Auto correlation

**4.3 Goldfeld Quant test:**

Homoscedasticity tested using Goldfeld Quant test.

Null hypothesis: Data satisfies the condition of Homoscedasticity

Alternate hypothesis: Data is not Homoscedastic



### 4.4.Linearity Assumption Test:

### 

### 

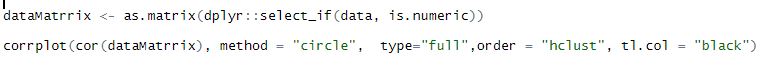
### 

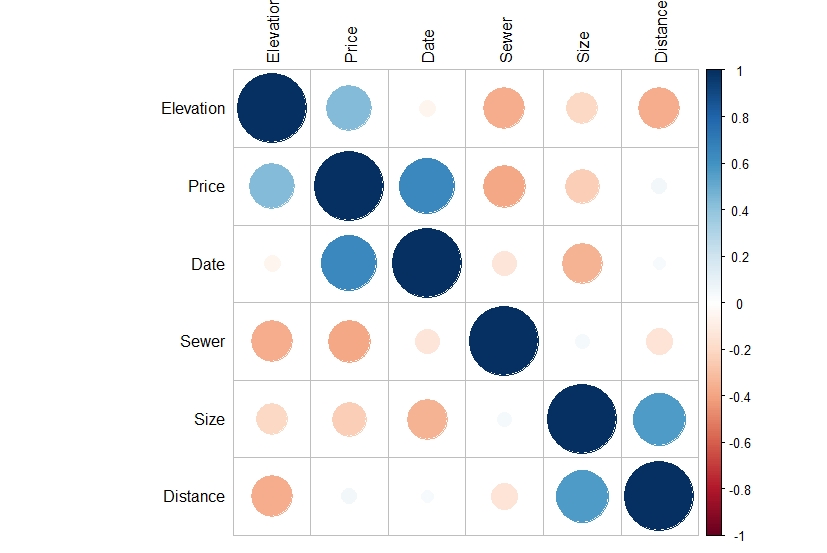
### 

### The variables does not seem to have defined linear relationship.But we will still go ahead with regression model and see the effects of such variables.

**4.5.Correlation:**

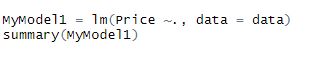
Let us take a look at how the variables are correlated with each other. We will omit the factor variables to do this because it will be difficult to compare categorical and numeric data correlation. Else we could try biserial.cor.But we will ignore that for now

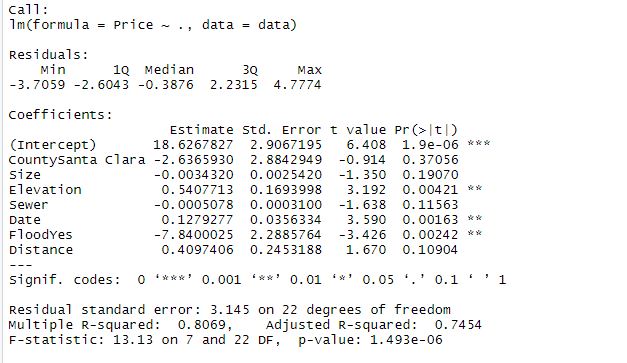




We can see that **Price** has a positive correlation with **Elevation** and **Date**.  
**Price** has a negative correlation with **Sewer**.  
**Price** has a minimal correlation with **Size** and **Distance**.

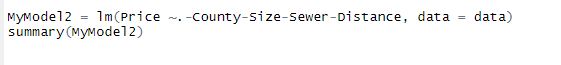
### 5. Model 1: All independent variables are considered to create a regression model

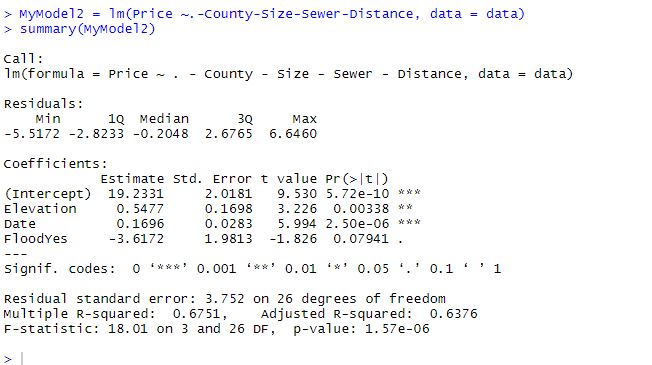




**Observation:** We have a very low p-value. Therefore, this is a valid model.  
When analysing the variable p-values, we can see that the variables **County, Size, Sewer and Distance** have a high p-value. Therefore, we will ignore these variables in our next regression model.

### 6. Model 2: Considering only those independent variable which has low P value in the new regression model



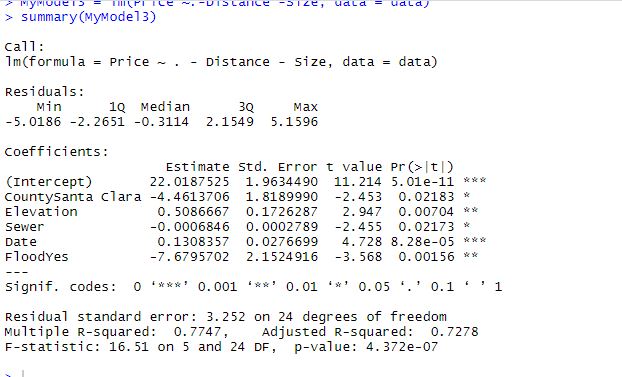


**Observation:**We again have a very low p-value. Therefore this is a valid model.  
One observation is the R-squared value has decreased compared to the previous model, therefore we will reject this model.

### 7. Model 3: Independent variable****s Size**** and ****Distance will not be considered to create a model****

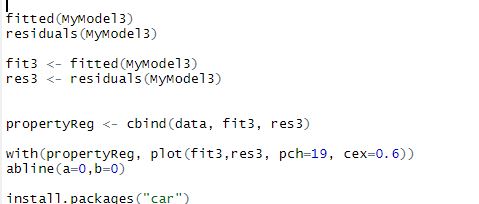
We will remove Distance and Size variables from our model as they are correlated as per the corrplot and this creates a problem of **Multicollinearity**

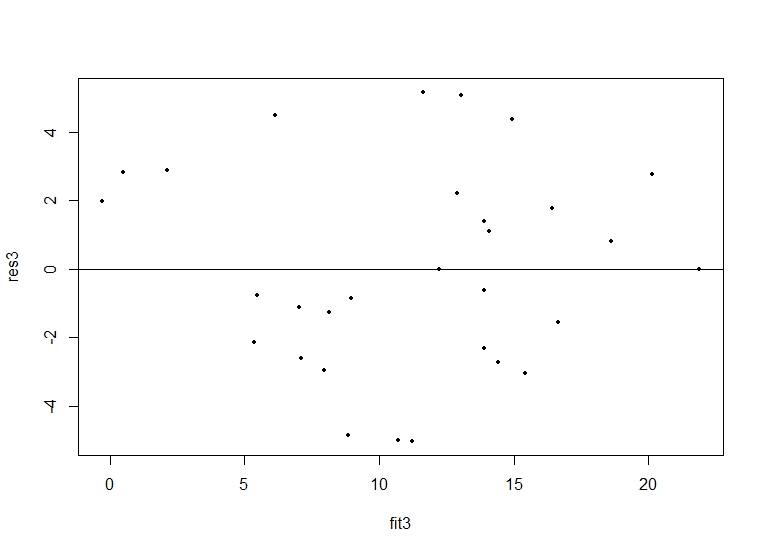




**Obsevation**:We again have a very low p-value and R-squared is higher than model 2. Therefore we will go with this model.

# **8. Residuals vs. Fits Plot for the appropriateness of the model.**



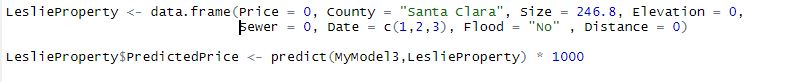


* The residuals "bounce randomly" around the 0 line. This suggests that the assumption that the relationship is linear is reasonable.
* The residuals roughly form a "horizontal band" around the 0 line. This suggests that the variances of the error terms are equal.
* No one residual "stands out" from the basic random pattern of residuals. This suggests that there are no outliers.

### 9. Let us predict the Leslie property value with the model 3:

**Problem Statement** : The Leslie property contained 246.8 acres and was located right on the San Francisco Bay. The land had been used for salt evaporation and had an elevation of exactly sea level. However, the property was diked so that the waters from the bay park were kept out.

Quantifying the above text into our final regression model we would be able to predict the property value as below:



### 10. Conclusion

The below table shows the Predicted prices of the Leslie property if purchased in the next 3 months.



