# Leslie Salt: Case Study

## Regression Analysis

Pratyush Yadav (MS18S004)

# Introduction

**Problem:** To determine a fair market value for the property based on collected data on 31 bayland properties that were sold in the last 10 years.

**Method and Solution Proposed:** Use of Regression Analysis to predict the price based on the given independent variables

**Variables:** Dependent Variable: Price (in $000 per acre)
Independent Variable :County (categorical), Size, Elevation (in ht), Sewer (distance), Date (backdated to purchase date, in months), Flood (categorical) and Distance (from Leslie Property).

# Getting Started: Scatter Plots and Correlation Analysis

```
leslie=read.csv("Leslie_Salt.csv",header=TRUE)
```

Getting information about the dataset using str function

```
str(leslie)
```

- It is important to inspect the variables in the dataset as to know what are their respective datatypes (numeric or int or factor).
- This will therefore aid in the proper understanding of what correlation actually gives. For eg: R treats each of the categorical variables as an integer value.
- There are different types of Correlation metrics used depending on the type of variables: Two continuous Variables - Pearson; One continous one categorical - Point Biserial; Two categorical - Cramer's V (based on Chi-square)
- In R, when you change the categorical variables from integer to factor (with levels), cor function returns an error!!

# Scatter Plots and Correlation Analysis

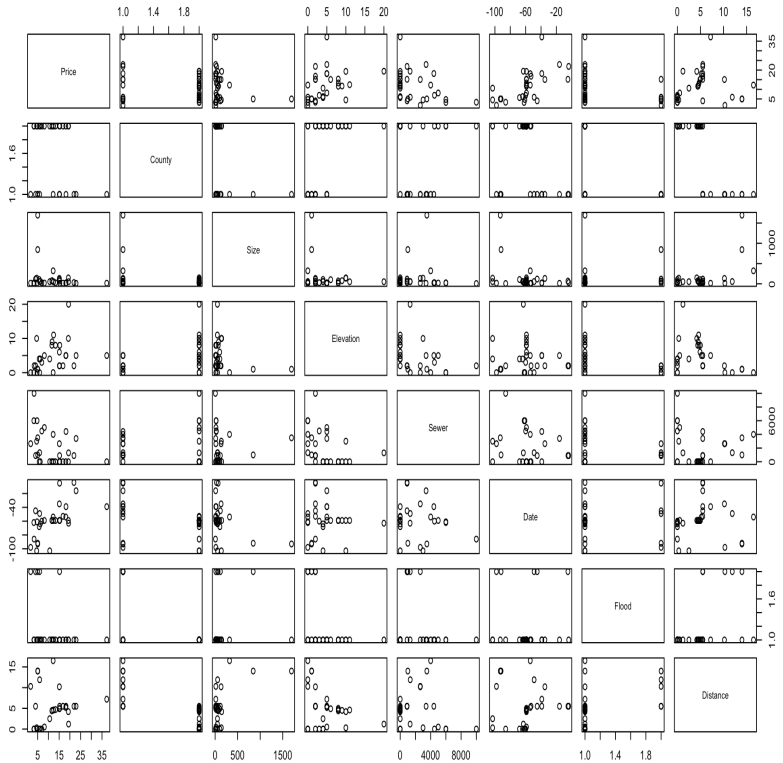
```
> cor(leslie)
              Price      County        Size   Elevation       Sewer        Date       Flood    Distance
Price      1.00000000 -0.18221231 -0.23973311  0.35184308 -0.39121198  0.59466427 -0.32309978  0.09331133
County    -0.18221231  1.00000000 -0.33944108  0.47517280 -0.05004423 -0.36983885 -0.55180357 -0.74220440
Size      -0.23973311 -0.33944108  1.00000000 -0.20945610  0.05338087 -0.34946290  0.10890203  0.55694587
Elevation  0.35184308  0.47517280 -0.20945610  1.00000000 -0.35940756 -0.05650853 -0.37308077 -0.36246039
Sewer     -0.39121198 -0.05004423  0.05338087 -0.35940756  1.00000000 -0.15149473 -0.11305464 -0.15865389
Date       0.59466427 -0.36983885 -0.34946290 -0.05650853 -0.15149473  1.00000000  0.01536084  0.04438251
Flood     -0.32309978 -0.55180357  0.10890203 -0.37308077 -0.11305464  0.01536084  1.00000000  0.42330840
Distance   0.09331133 -0.74220440  0.55694587 -0.36246039 -0.15865389  0.04438251  0.42330840  1.00000000
```

Getting the correlation matrix

`cor(leslie)`

Plotting the pairwise scatterplots

`pairs(leslie)]`

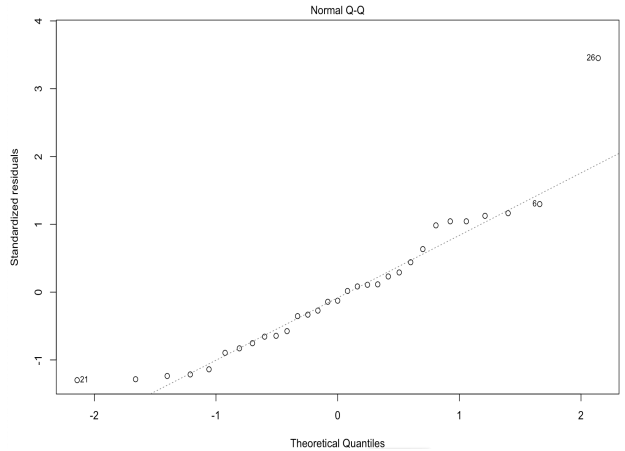


# Regression Models

```
model1=lm(Price~.,data=leslie)
summary(model1)
library(car)
vif(model1)
```

```
Call:
lm(formula = Price ~ ., data = leslie)

Residuals:
   Min     1Q Median     3Q    Max
-5.169 -2.957 -0.256  2.070 13.031

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.364e+01  3.829e+00   6.174 2.68e-06 ***
CountySanta Clara -8.789e+00  3.652e+00  -2.407 0.024532 *
Size              -6.043e-03  3.501e-03  -1.726 0.097702 .
Elevation          5.193e-01  2.386e-01   2.177 0.040030 *
Sewer             -9.573e-04  4.169e-04  -2.296 0.031126 *
Date               8.508e-02  4.865e-02   1.749 0.093646 .
FloodYes          -1.202e+01  2.989e+00  -4.020 0.000536 ***
Distance           1.858e-01  3.395e-01   0.547 0.589386
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.431 on 23 degrees of freedom
Multiple R-squared:  0.747,     Adjusted R-squared:   0.67
F-statistic: 9.703 on 7 and 23 DF,  p-value: 1.351e-05
```

```
> vif(model1)
   County      Size Elevation     Sewer      Date     Flood  Distance
4.995597  2.003925  1.649759  1.635122  2.174889  1.907942  3.623612
```




- Except Distance, Size and Date, every variable is important and hence we intend to remove one of these variables in the next model
- The qq-plot looks pretty good in terms of closeness to normal except for point 26, which is an outlier - what happens when we remove it?

# Regression Models-contd

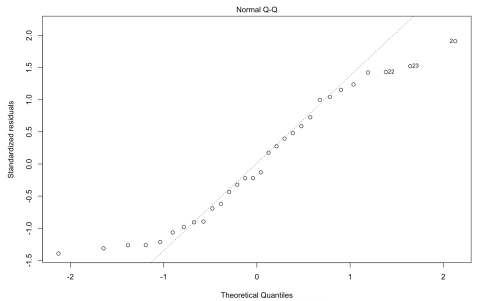
```
model2=lm(Price~.-Distance,data=leslie)
summary(model2)
```

Removing Size and Date led to the decrease in Adjusted R-Square value, which indicates that they are rather significant and they shouldn't be removed, hence only distance was eliminated

```
Call:
lm(formula = Price ~ . - Distance, data = leslie)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6249 -2.8459 -0.5266  1.9702 12.6368

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.521e+01  2.514e+00  10.025 4.68e-10 ***
CountySanta Clara -1.005e+01  2.789e+00  -3.604 0.001423 **
Size              -5.425e-03  3.265e-03  -1.662 0.109585
Elevation          4.995e-01  2.324e-01   2.150 0.041869 *
Sewer             -1.054e-03  3.724e-04  -2.829 0.009272 **
Date               7.842e-02  4.641e-02   1.690 0.104029
FloodYes          -1.219e+01  2.927e+00  -4.166 0.000347 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.366 on 24 degrees of freedom
Multiple R-squared:  0.7437,    Adjusted R-squared:  0.6797
F-statistic: 11.61 on 6 and 24 DF,  p-value: 4.257e-06
```

```
model3=lm(Price~.,data=leslie[-26,])
summary(model3)
```

```
Call:
lm(formula = Price ~ ., data = leslie[-26, ])

Residuals:
    Min      1Q  Median      3Q     Max
-3.7059 -2.6043 -0.3876  2.2315  4.7774

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       18.6267827  2.9067195   6.408  1.9e-06 ***
CountySanta Clara -2.6365930  2.8842949  -0.914  0.37056
Size              -0.0034320  0.0025420  -1.350  0.19070
Elevation          0.5407713  0.1693998   3.192  0.00421 **
Sewer             -0.0005078  0.0003100  -1.638  0.11563
Date               0.1279277  0.0356334   3.590  0.00163 **
FloodYes          -7.8400025  2.2885764  -3.426  0.00242 **
Distance           0.4097406  0.2453188   1.670  0.10904
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.145 on 22 degrees of freedom
Multiple R-squared:  0.8069,    Adjusted R-squared:  0.7454
F-statistic: 13.13 on 7 and 22 DF,  p-value: 1.493e-06
```




The QQ-plot after removing the outlier deviates a lot from the normal curve - normality assumption violated!!!

# Regression Models

```
leslie$Price1=log(leslie$Price)
model4=lm(Price1~.-Price,data=leslie)
summary(model4)
```

```
Call:
lm(formula = Price1 ~ . - Price, data = leslie)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41605 -0.22833  0.01037  0.22662  0.63418

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.099e+00  2.815e-01  11.006 1.22e-10 ***
CountySanta Clara -1.596e-01  2.685e-01  -0.594 0.558013
Size              -2.578e-04  2.574e-04  -1.002 0.327001
Elevation          5.053e-02  1.754e-02   2.880 0.008448 **
Sewer             -8.338e-05  3.066e-05  -2.720 0.012214 *
Date               1.479e-02  3.577e-03   4.135 0.000403 ***
FloodYes          -9.819e-01  2.198e-01  -4.468 0.000175 ***
Distance           4.889e-02  2.496e-02   1.958 0.062407 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3258 on 23 degrees of freedom
Multiple R-squared:  0.8416,    Adjusted R-squared:  0.7934
F-statistic: 17.46 on 7 and 23 DF,  p-value: 8.112e-08
```

County, Size and Distance did not end up being significant, hence we can afford to remove one of these variables to get a better model fit. Since the p-value of County is the largest, I remove it first.

```
model5=lm(Price1~.-Price-County,data=leslie)
summary(model5)
```

```
Call:
lm(formula = Price1 ~ . - Price - County, data = leslie)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37796 -0.22920 -0.01371  0.20334  0.68359

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.006e+00  2.310e-01  13.009 2.31e-12 ***
Size        -2.256e-04  2.482e-04  -0.909  0.37240
Elevation    4.992e-02  1.727e-02   2.890  0.00806 **
Sewer       -7.653e-05  2.802e-05  -2.731  0.01164 *
Date         1.614e-02  2.729e-03   5.914 4.22e-06 ***
FloodYes    -9.154e-01  1.866e-01  -4.906 5.27e-05 ***
Distance     5.826e-02  1.908e-02   3.053  0.00547 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3214 on 24 degrees of freedom
Multiple R-squared:  0.8392,    Adjusted R-squared:  0.799
F-statistic: 20.87 on 6 and 24 DF,  p-value: 1.978e-08
```

The Adjusted R-square has increased, which means we have removed a value which might not be contributing much to the model. Moreover, removal of County has led to **Distance becoming significant**

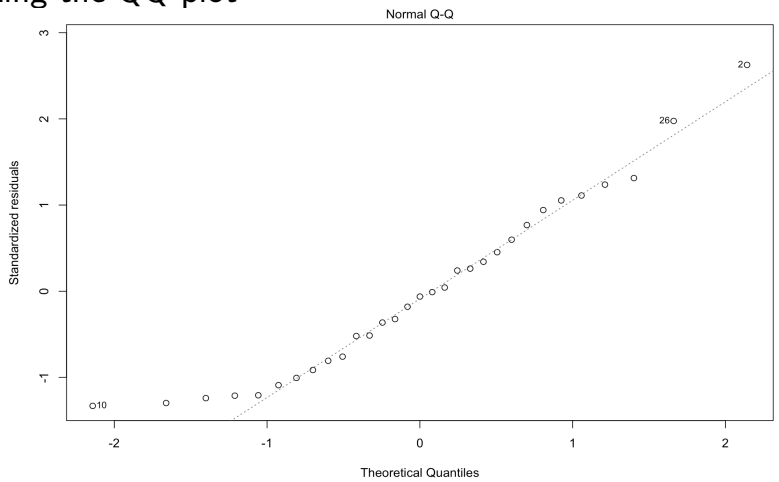
# Final Model

```
leslie$Price1=log(leslie$Price)
model6=lm(Price1~.-Price-County-Size,data=leslie)
summary(model6)
```

```
Call:
lm(formula = Price1 ~ . - Price - County - Size, data = leslie)

Residuals:
     Min       1Q   Median       3Q      Max
-0.38511 -0.25256 -0.01794  0.20994  0.72640

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.089e+00  2.115e-01  14.603 9.60e-14 ***
Elevation    5.048e-02  1.720e-02   2.934  0.00707 **
Sewer       -7.859e-05  2.783e-05  -2.824  0.00919 **
Date         1.724e-02  2.435e-03   7.080 2.02e-07 ***
FloodYes    -8.835e-01  1.826e-01  -4.838 5.66e-05 ***
Distance     4.784e-02  1.521e-02   3.147  0.00424 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3203 on 25 degrees of freedom
Multiple R-squared:  0.8336,    Adjusted R-squared:  0.8003
F-statistic: 25.05 on 5 and 25 DF,  p-value: 5.474e-09
```

Removing Size after County gives the best model with a high Adjusted R Square. All the other variables are significant

Visualizing the QQ-plot




The residuals almost follow the straight line closely. To solidify this conclusion, we perform the Shapiro wilk test

```
shapiro.test(model6$residuals)
```

```
        Shapiro-Wilk normality test

data:  model5$residuals
W = 0.94417, p-value = 0.1077
```

# Conclusion

- The final model for Leslie Property Price Prediction used log transformation on the dependent variable, i.e. log(Price) and the independent variables which turned out to be significant included Elevation, Sewer,Date,Flood and Distance.

$$log(Price) = \beta_0 + \beta_1 \times Elevation + \beta_2 \times Sewer + \beta_3 \times Date + \beta_4 \times Flood + \beta_5 \times Distance$$

- There was no significant multicollinearity found in the data
- Shapiro-Wilk test indicated that the residuals are normally distributed ($W = 0.94417$, $p - value = 0.1077$)
- Adjusted R-Square was used as the metric for model-fit as it ensures the proper balance between adequate number of variables and inclusion of significant variables. The value for the final model was found to be 0.8003.
- Best subset selection or forward and backward selection methods can also be used in order to find the final model rather than building models subsequently. In R, this can be implemented using stepAIC function and it uses AIC (Akaike Information Criterion) as a metric for adding or dropping variables in the model.
- Note: This model was the chosen out of all the models tried, however it isn't the best model per se.
- There may be other considerations which haven't been taken into account like Interaction Effects