INTRODUCTION

Dyer et al. created a biodiversity-based indicator that may be used for large-scale environmental assessments, utilizing prospective shale gas extraction sites in the United Kingdom as a case study. They discovered that the proposed shale gas sites were often located in areas of relatively low species richness nationally by analyzing data on the distribution of 855 species across the United Kingdom. However, evaluations at smaller sizes did identify certain regions with significant local biodiversity value within the sites. The authors provide a technique for locating crucial biodiversity protection zones inside extraction sites on a local and regional level. This biodiversity indicator could assist in environmental assessments and planning decisions for significant developments like shale gas, assisting in the balancing of environmental, social, and economic objectives. The study shows the importance of biodiversity data for creating indicators that can estimate the possible effects of industrial activity on the ecosystem.

DATA EXPLORATION

A dataset of proportional species has been sent to us. It has 5281 rows of data points, 17 columns of features, 11 taxonomic groups, and additional columns for easting, northing, dominant class, ecological status, and time.

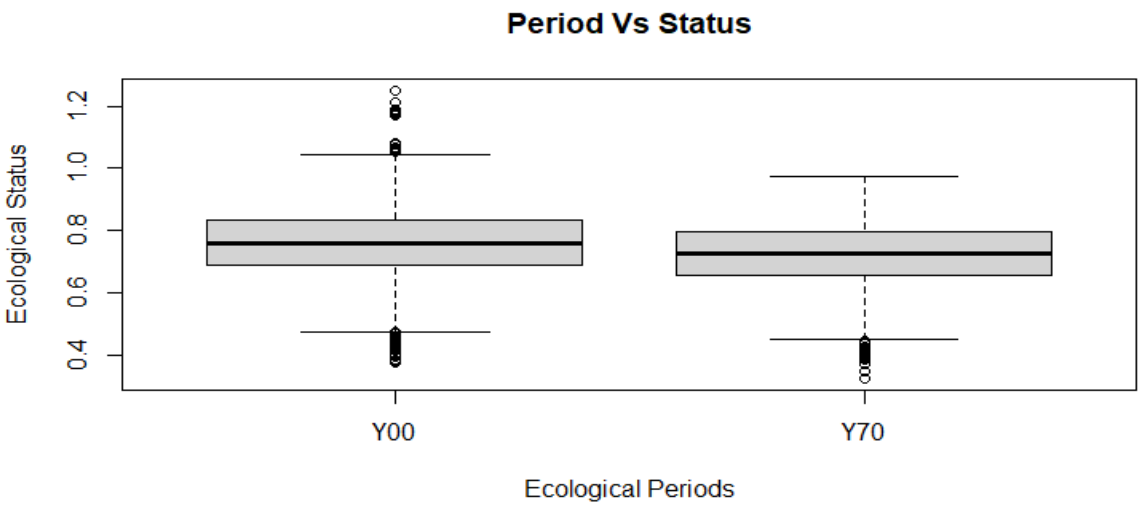
From the 11 taxonomic groupings, we must choose seven to analyse: bees, birds, bryophytes, butterflies, isopods, ladybirds, and macromoths.

Fig: 1 Summary of the Dataset

```
> summary(Proj_data)
   Location             Bees              Bird            Bryophytes      Butterflies
 Length:5280        Min.   :0.03065   Min.   :0.2415   Min.   :0.3941   Min.   :0.3167
 Class :character   1st Qu.:0.35079   1st Qu.:0.8462   1st Qu.:0.6886   1st Qu.:0.7926
 Mode  :character   Median :0.58869   Median :0.9038   Median :0.7993   Median :0.8863
                    Mean   :0.60502   Mean   :0.8872   Mean   :0.7866   Mean   :0.8746
                    3rd Qu.:0.81663   3rd Qu.:0.9570   3rd Qu.:0.8855   3rd Qu.:0.9677
                    Max.   :3.30986   Max.   :1.1720   Max.   :1.1746   Max.   :1.3944

    Carabids         Hoverflies        Isopods          Ladybirds        Macromoths
 Min.   :0.01153   Min.   :0.1235   Min.   :0.04622   Min.   :0.0614   Min.   :0.08947
 1st Qu.:0.47539   1st Qu.:0.5696   1st Qu.:0.39165   1st Qu.:0.4545   1st Qu.:0.78555
 Median :0.63553   Median :0.6957   Median :0.53936   Median :0.6395   Median :0.87667
 Mean   :0.60706   Mean   :0.6795   Mean   :0.54995   Mean   :0.6140   Mean   :0.84927
 3rd Qu.:0.76161   3rd Qu.:0.8063   3rd Qu.:0.71623   3rd Qu.:0.7972   3rd Qu.:0.94152
 Max.   :1.19977   Max.   :1.1453   Max.   :1.25773   Max.   :1.8400   Max.   :1.26045

 Grasshoppers_._Crickets Vascular_plants     Easting          Northing        dominantLandClass
 Min.   :0.0708          Min.   :0.4179   Min.   : 50000   Min.   :  10000   3e      : 346
 1st Qu.:0.4876          1st Qu.:0.7213   1st Qu.:260000   1st Qu.: 220000   2e      : 324
 Median :0.6250          Median :0.7912   Median :340000   Median : 390000   10e     : 316
 Mean   :0.6289          Mean   :0.7869   Mean   :352886   Mean   : 452136   1e      : 260
 3rd Qu.:0.7934          3rd Qu.:0.8551   3rd Qu.:440000   3rd Qu.: 670000   25s     : 214
 Max.   :1.5938          Max.   :1.2023   Max.   :650000   Max.   :1210000   22s     : 212
                                                                             (Other):3608

 ecologicalStatus period
 Min.   :0.3538   Y00:2640
 1st Qu.:0.6518   Y70:2640
 Median :0.7191
 Mean   :0.7154
 3rd Qu.:0.7899
 Max.   :1.1071

>
```

The table depicts the distribution of ten distinct insect and plant species in 5280 different places. Bees are the most prevalent species, and vascular plants are the least prevalent, with species abundance varying from place to place. Bees, birds, and butterflies are most prevalent in forests, and vascular plants are most prevalent in grasslands. The number of both insects and plants is connected with the dominant land class. Bees, birds, and butterflies are most prevalent in areas with good ecological status, while vascular plants are most prevalent in areas with poor ecological status. The ecological state of the place also impacts the amount of insects and plants.

Fig: 2 Box plot Plotted Between the Ecological Period vs Ecological Status



The graph depicts the relationship between the ecological period and the ecological status, and the trends show that the ecological period lengthens as the ecological status shortens, according to the above bar plot. Additionally, we can see from the graph that there is a great deal of variety.

Fig 3: Skew, Standard Deviation and Mean of the & Taxonomic Groups, SD, and Mean Table

```
> summary_stats
           Bees       Bird Bryophytes Butterflies    Isopods  Ladybirds Macromoths
mean 0.6050238  0.8871739  0.7865969   0.8745706 0.54994958 0.61403361  0.8492665
sd   0.3105858  0.1065161  0.1317877   0.1396994 0.21545995 0.26603380  0.1406266
skew 0.9580630 -1.5070945 -0.1982652  -0.3599746 0.04789114 0.03409944 -1.1385102
```

Statistics on six different insect species are summarised. In comparison to other insect populations, bee populations appear to have a higher degree of size variability because they have the lowest mean value and the highest standard deviation. Compared to other insect populations, ladybird populations appear to be more stable in size since they have the highest mean value and the lowest standard deviation. Various variables, such as the diverse habitats that bees and ladybirds utilise, may be to blame for these variations in population distribution.

HYPOTHESIS TESTING

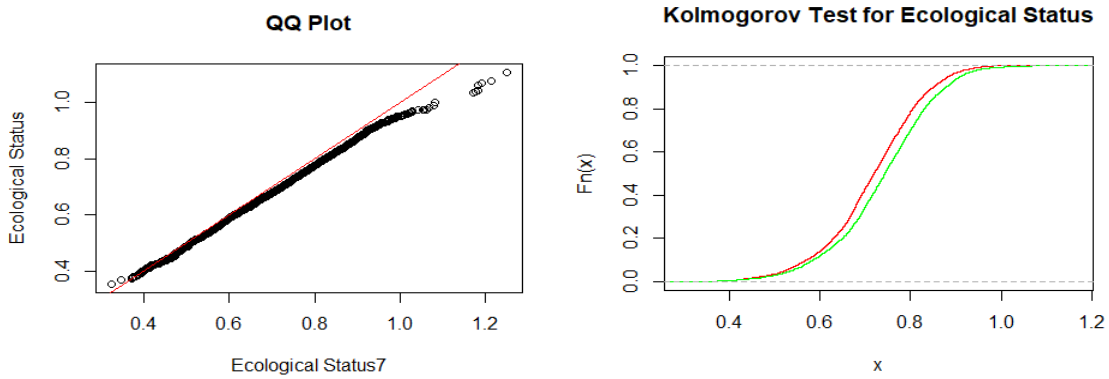
To get the Statistical Importance of different Sample test variants.

Fig 4: Output of the One Sample T-test

```
> t.test(BD7_change,mu=0)  # t test with H0: mu=0

        One Sample t-test

data:  BD7_change
t = 26.209, df = 2639, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.03421060 0.03974365
sample estimates:
 mean of x
0.03697713
```

The One sample test is performed H0 which is Null Hypothesis. From the above result shown which tells df is having the 2639 participants with a T value of 26.209 and P value <2.2e-16.

Fig 5: QQ Plot and Kolmogorov Test Curve Plot



The QQ plot shows that the data points lying on the regression line do not fall perfectly and are not distributed properly, and by observing the tails, we can conclude that the data is left-skewed in this particular case as the lower tail has more data points than the upper tail.

Simple Linear Regression

Fig 6: Summary of Simple Linear Regression

```
> summary(lin_mod)

Call:
lm(formula = Proj_data_MA334$eco_status_7 ~ Proj_data_MA334$ecologicalStatus)

Residuals:
      Min        1Q    Median        3Q       Max
-0.130241 -0.026160 -0.001962  0.023994  0.194646

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   0.037655   0.003818   9.863   <2e-16 ***
Proj_data_MA334$ecologicalStatus 0.979136   0.005277 185.546   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04143 on 5278 degrees of freedom
Multiple R-squared:  0.8671,    Adjusted R-squared:  0.867
F-statistic: 3.443e+04 on 1 and 5278 DF,  p-value: < 2.2e-16
```

In the Simple Linear Regression, we have taken Proj_data_MA334$eco_status_7 is Response or Dependent Variable and Proj_data_MA334$ecologicalStatus predictor or independent variable.

According to the model, a rise in ecological Status of 1 unit is indicated by an intercept of 0.037655 and a slope coefficient of 0.979136.

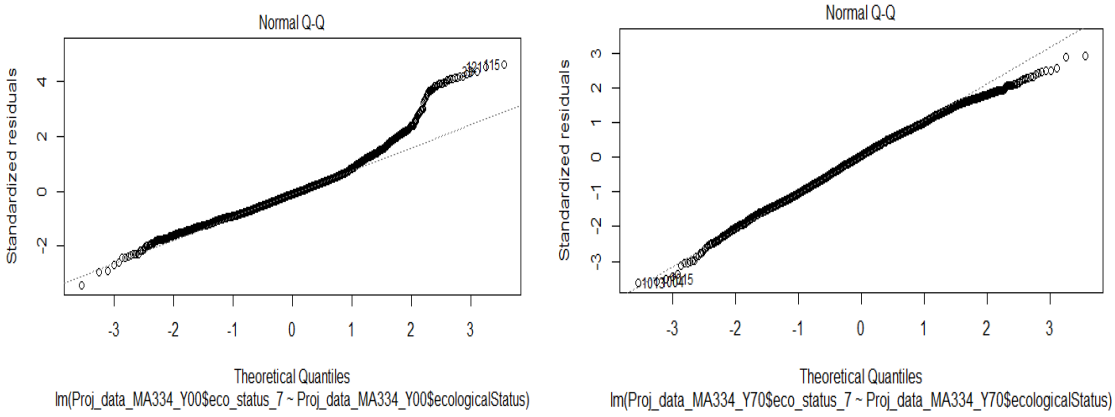
The model has very high Multiple R-Squared Values which is 0.8671 which says that ecological Status explain 86.71% of the variance in eco_status_7.

The ecological Status coefficient is extremely statistically significant, according to the t-test p<2.2e-16

Therefore, we can say that this is a strong Model.

For Each period:

Fig 7: QQ Plot



Multiple Linear Regression

Fig 8: Summary of Multiple Linear Regression

```
> summary(lmMod_train)

Call:
lm(formula = eco_status_4 ~ ., data = trainingData[c(eco_selected_names,
    "eco_status_4")], na.action = na.omit, y = TRUE)

Residuals:
     Min       1Q   Median       3Q      Max
-0.42972 -0.05809 -0.00052  0.06949  0.31761

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.145798   0.017821   8.181 3.68e-16 ***
Bees         0.091411   0.006338  14.422  < 2e-16 ***
Bird         0.227706   0.019632  11.599  < 2e-16 ***
Bryophytes   0.119995   0.012499   9.600  < 2e-16 ***
Butterflies -0.048728   0.014346  -3.397 0.000689 ***
Isopods      0.231061   0.007514  30.752  < 2e-16 ***
Ladybirds    0.082175   0.008386   9.799  < 2e-16 ***
Macromoths   0.051657   0.016576   3.116 0.001844 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1011 on 4216 degrees of freedom
Multiple R-squared:  0.4198,    Adjusted R-squared:  0.4188
F-statistic: 435.7 on 7 and 4216 DF,  p-value: < 2.2e-16
```

In Multiple Linear Regression Model, the Response or Dependent Variable is eco_status_4 and the predictor or independent variable with all 7 taxonomic group.

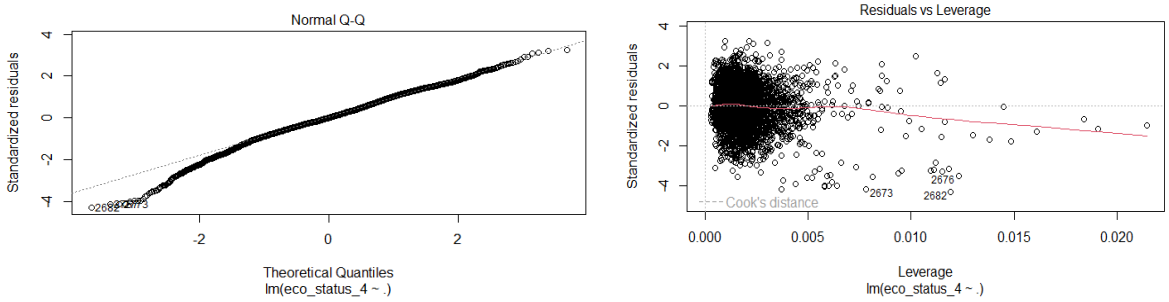
From the Above the image of Summary of Multiple Linear Regression Model which says that the residual values (0.105) is low and R-Squared Value of 0.424 which means indicates that the model is able to predict the eco_status_4.The coefficient values indicates that all variable are positively correlated with the eco_status_4 variable except the Butterflies, which means that an increase in the value of any of positive variables is associated with increase in the value of eco_status_4.To conclude, We can say that significance code is statistically significant at 0.05 level which suggest that model can be for better understanding.

Fig 9: Correlation Between Train and Test

```
> cor(lmMod_train$fitted.values,lmMod_train$y) # cor training data
[1] 0.6534089
```

The correlation coefficient is 0.6534089, indicating that the two variables have a favourable relationship. This indicates that although the two variables have a good relationship, it is not perfect.

To make it more clear and clarify We can see in the plot



Now, we are going to Feature Engineering in which will not consider one variable because the one is not statistically significant for the model.

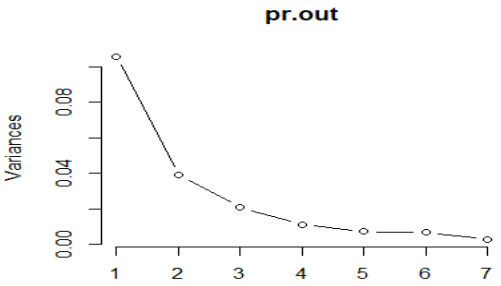
```
> summary (lmModdel_reduced_train2)

Call:
lm(formula = eco_status_4 ~ ., data = trainingData[c(eco_selected_names
[-7],
    "eco_status_4", "period", "Northing")], na.action = na.omit,
    y = TRUE)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33904 -0.05632  0.00010  0.05700  0.38243

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.028e-01  1.633e-02   6.296 3.37e-10 ***
Bees         1.438e-01  5.437e-03  26.452  < 2e-16 ***
Bird         2.373e-01  1.564e-02  15.171  < 2e-16 ***
Bryophytes   1.214e-01  1.076e-02  11.288  < 2e-16 ***
Butterflies  9.747e-02  1.167e-02   8.351  < 2e-16 ***
Isopods      2.787e-02  7.902e-03   3.527 0.000424 ***
Ladybirds    7.345e-02  6.913e-03  10.625  < 2e-16 ***
periodY70    1.458e-01  3.673e-03  39.693  < 2e-16 ***
Northing    -8.487e-08  5.707e-09 -14.871  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08462 on 4215 degrees of freedom
Multiple R-squared:  0.5935,    Adjusted R-squared:  0.5927
F-statistic: 769.2 on 8 and 4215 DF,  p-value: < 2.2e-16
```

AIC

```
> AICValues
[1] -7364.273 -8865.039
```

AIC value is one of the measures to check that the model is fit or not. So, in the above image we have two values -7364.273 and -8865.039. -7364.273 AIC value is a better fir for the data then the model with the AIC value of -8865.039.

OPEN ANALYSIS




```
> pr.out$center  # gives the mean corrections the "cen
ters"
Vascular_plants           Bees     Butterflies
    -0.06110240     0.20545688      0.11458279
      Carabids           Bird      Hoverflies
    -0.19468170     0.03264483     -0.08242952
     Bryophytes
      0.00549104
```

The mean correction is shown to be to the left of the decimal point by the negative values, and to the right by the positive numbers. In the case of vascular plants, the average correction is -0.0611024, which indicates that the centres of these plants are typically 0.0611024 to the left of the decimal point.

Conclusion

Bees, birds, and butterflies are more common in forests with high ecological value, according to a survey that examined species distribution data across the UK. Significant relationships between the number of species and ecological status have been shown by statistical analysis. The data indicates protecting areas of good ecological status is key to maintaining biodiversity of vulnerable species. Detailed species data can inform ecological health indicators to support sustainable environmental planning. Further analysis on individual species and additional variables like climate change could provide further insights.

References:

- Developing a biodiversity-based indicator for large-scale environmental assessment: a case study of proposed shale gas extraction sites in Britain Robert James Dyer1 , Simon Gillings2 , Richard F. Pywell1 , Richard Fox3 , David B. Roy1 and Tom H. Oliver4 *