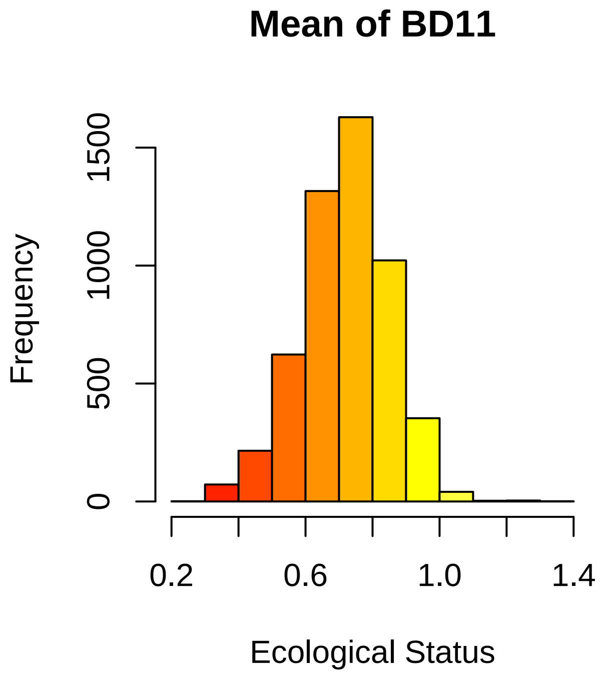
Proportional Species Richness

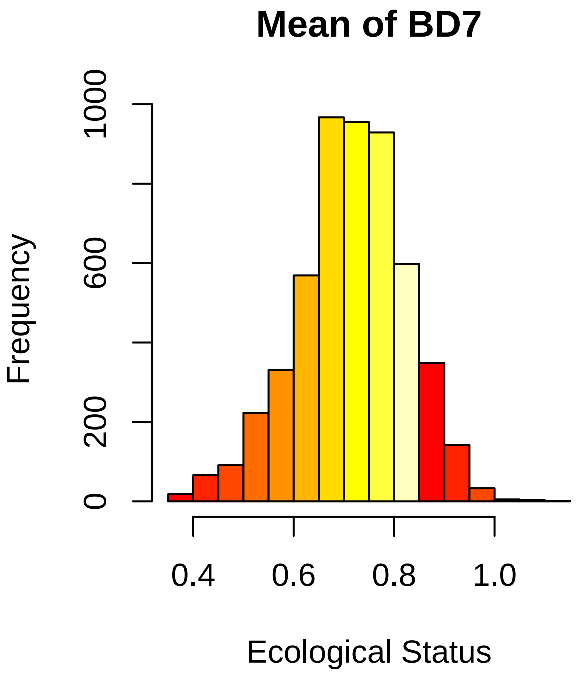
Introduction

Bees, Macromoths, Ladybirds, Butterflies, Vascular_plants, Hoverflies, and Grasshoppers_._Crickets were the seven features we focused on during exploratory data analysis. Our goal was to find relations and trends that might be useful for further research by examining the possible connections between these characteristics. We were further provided with the data about Location, Eastings, Northings, DominantLandClass , ecological status, period (years between 1970 and 1990 named as period Y90 and span of 2000 to 2013 into period Y00).

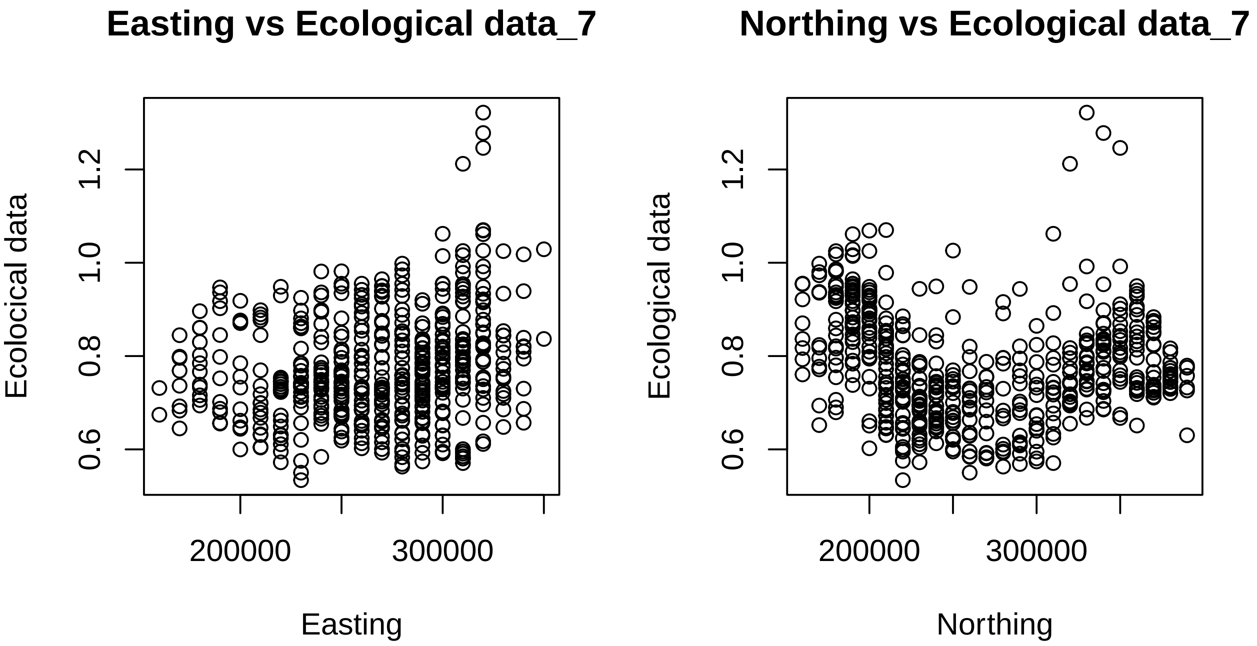
Exploratory Data Analysis (EDA)

The study found some intriguing results in our data analysis. The mean of the ecological status for all the taxonomy groups (BD11) to that of the mean of the seven features (BD7) that were allocated were both almost similar.(See below plot) However, when we took the mean of each feature separately, we found a significant change with the Bees (0.58)and Macromoths (0.86). Furthermore, the number of Ladybirds, Macromoths and butterflies was found to have a moderate positive correlation. This suggests that the when the Ladybirds population improves, butterflies and Macromoths population typically improves as well. This holds true for Hoverflies, and Grasshoppers_._Crickets too. This link is significant because it implies that these species may need one another for survival.





On the other hand, we took a sample of data for the wales region and there is a positive correlation (0.21) between the BD7 and the Easting, which meant that as we go towards the east the ecological status also tends to increase. However, towards the North we have a negative correlation (-0.11) which suggest that as we go towards the North the ecology decreases.



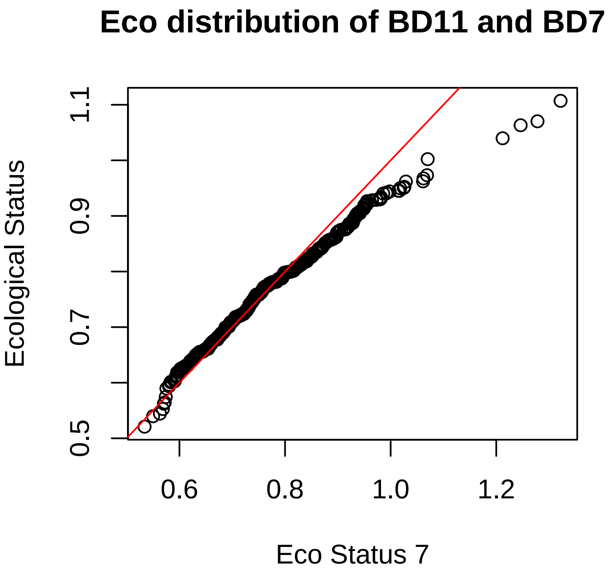
Hypothesis Testing

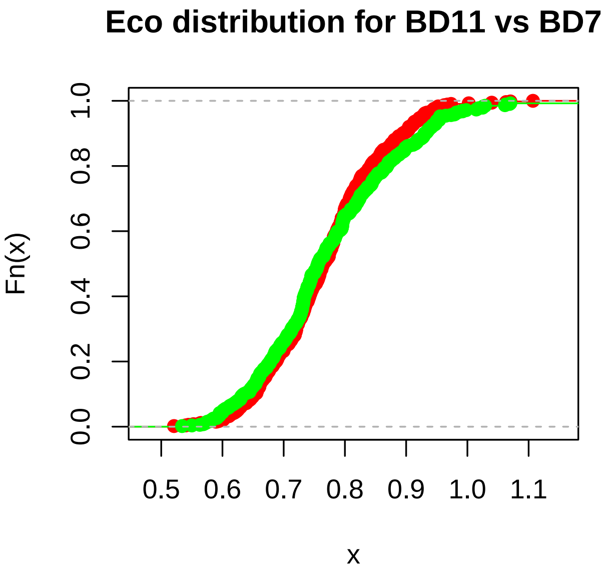
Based on the results of the one-sample t-test, there is strong evidence to suggest that the true mean of the population (BD7) is not equal to 0. The calculated t-value was 5.5898 which is significant at the chosen significance level, and the 95 percent confidence interval does not include 0. This implies that there is a statistically significant difference between the sample mean and the hypothesized population mean of 0.

```
##
##  One Sample t-test
##
## data:  BD7_change
## t = 5.5898, df = 262, p-value = 5.701e-08
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.02384851 0.04978780
## sample estimates:
##  mean of x
## 0.03681815
```

We have also performed the Asymptotic two-sample Kolmogorov-Smirnov test, where we used the BD7 population and the BD11 population for their ecological status analysis. The text statistic (D) was 0.068441, and the p-value that was associated to it was 0.1701. The alternative hypothesis was two – sided and also the p-value being higher than the significance level led to no significant difference between both the distributions.

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  Proj_data_MA334$eco_status_7 and Proj_data_MA334$ecologicalStatus
## D = 0.068441, p-value = 0.1701
## alternative hypothesis: two-sided
```





Simple linear regression

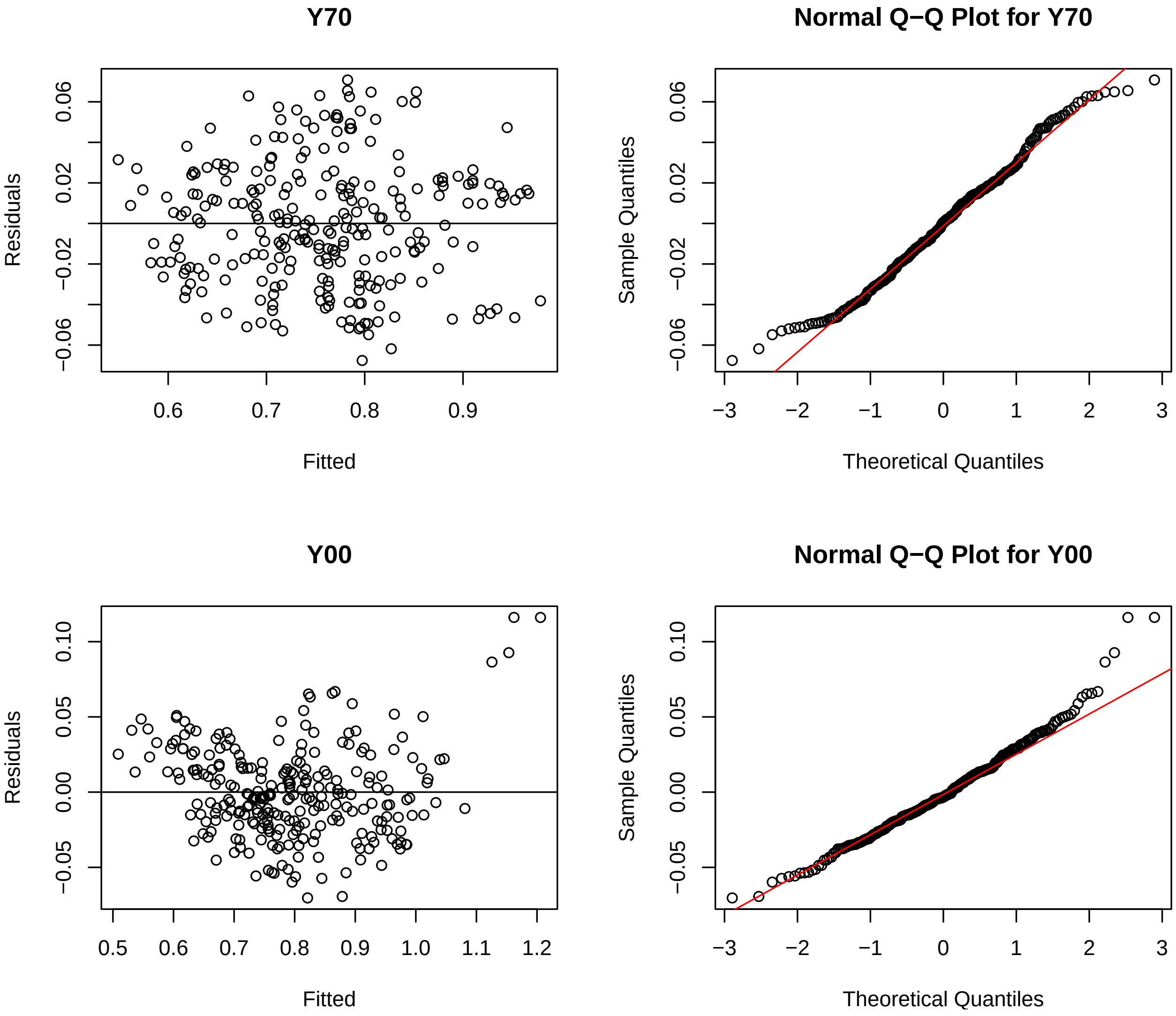
A basic linear regression model with one predictor variable (Bd11) and one response variable (BD7) for the two different time periods in Wales and is displayed in summary. The explanatory variable's estimated Intercept and slope coefficient are displayed in the coefficient table. Both coefficients have very modest p-values (less than 0.001), suggesting they are likely to be non-zero. The R-squared score for Y70 was 0.9022 indicates that the predictor variable accounts for 90.22% of the variance in the responder variable. The residual standard error was 0.0307. On the other hand Y00 resulted in R-squared score of 0.9441 , indicating the predicator variable accounts for 94.41% of variance in the responder variable. The residual standard error was 0.0303 for Y00 which is quite similar tohat of the what it was for Y70. The coefficient was -0.17073 in the Y70 and - 0.11068 in the Y00 which indicates a negative relation between the predictor variable and the response variable. There is substantial evidence that the predictor variable significantly contributes to explaining the variance in the response variable, as indicated by the F-huge statistic's value of 5.335e+04 and low p-value (0.001).

The above findings also suggests that the relationship was stable for a relatively good time and then gets negative over time.

```
##
## Call:
## lm(formula = Bd7 ~ Bd11, data = df_filtered_Y70)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.067589 -0.022220  0.000289  0.019779  0.070752
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.17073    0.01898  -8.995   <2e-16 ***
## Bd11         1.19468    0.02434  49.081   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03037 on 261 degrees of freedom
## Multiple R-squared:  0.9022, Adjusted R-squared:  0.9019
## F-statistic:  2409 on 1 and 261 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = Bd7 ~ Bd11, data = df_filtered_Y00)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.070318 -0.019621 -0.003274  0.016462  0.116104
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.11068    0.01374  -8.056 2.83e-14 ***
## Bd11         1.18910    0.01791  66.404  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03033 on 261 degrees of freedom
## Multiple R-squared:  0.9441, Adjusted R-squared:  0.9439
## F-statistic:  4409 on 1 and 261 DF,  p-value: < 2.2e-16
```

The below plots gives us information



Multiple Linear Regression

Bees, Macromoths, Ladybirds, Butterflies, Vascular_plants, Hoverflies, and Grasshoppers_._Crickets are all included in the table of calculated coefficients. When all predictor variables are set to zero, the Intercept is 0.291024.

When we have removed on significant variable, herein Butterflies, then the Ladybirds, Vascular_plants, Hoverflies, and Grasshoppers_._Crickets all will have positive coefficients, indicating that rising populations of these organisms correlate with rising response variable levels. For instance, a one-unit increase in the Ladybirds population relates to a 0.089974 unit rise in the response variable. However, when the Macromoths and Bees population increase, the target variable declines, as seen by the negative coefficient.

The multiple R-squared for the model is 0.5153, which indicates that the predictor variables account for 51.53 percent of the variance in the response variable. Taking into account the total number of predictors, the corrected R-squared value is 0.5148. With a p-value of 2.2e-16, the F-statistic is statistically significant, suggesting that one or more of the predictor variables significantly contributes to the model. Unaccounted-for volatility in the response variable is represented by the residual standard error (0.07173), which is the standard deviation of the residuals.

```
##
## Call:
## lm(formula = BD4 ~ Bees + Macromoths + Ladybirds + Vascular_plants +
##     Hoverflies + Grasshoppers_._Crickets, data = Main_Proj_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.233889 -0.045489  0.003235  0.049224  0.214401
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.255622   0.009536  26.806  < 2e-16 ***
## Bees                    -0.026395   0.003940  -6.699 2.32e-11 ***
## Macromoths              -0.020861   0.008767  -2.380   0.0174 *
## Ladybirds                0.089974   0.004733  19.009  < 2e-16 ***
## Vascular_plants          0.296969   0.010800  27.496  < 2e-16 ***
## Hoverflies               0.244329   0.006771  36.084  < 2e-16 ***
## Grasshoppers_._Crickets  0.048990   0.005574   8.789  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07173 on 5273 degrees of freedom
## Multiple R-squared:  0.5153, Adjusted R-squared:  0.5148
## F-statistic: 934.4 on 6 and 5273 DF,  p-value: < 2.2e-16
```

The report in (figure 4) shows the outcomes of a stepwise selection procedure in linear regression using the Akaike Information Criterion (AIC) as the model selection criterion. The model first runs the entire model, which includes all predictor variables (Bees, Macromoths, Ladybirds, Butterflies, Vascular Plants, Hoverflies, Grasshoppers_.Crickets), and then goes through a stepwise selection procedure, deleting one variable at a time, until the AIC does not improve significantly. The output shows the changes in the AIC, sum of squares (sum of squares of residuals), and residual sum of squares (RSS) after removing each variable from the entire model. The AIC values show the quality of model fit, with lower values suggesting better fit. The findings show that all predictor variables are present in the optimal model for BD4 prediction (with the lowest AIC).

```
## Start:  AIC=-27900.9
## BD4 ~ Bees + Macromoths + Ladybirds + Butterflies + Vascular_plants +
##     Hoverflies + Grasshoppers_._Crickets
##
```

```
##                            Df Sum of Sq    RSS    AIC
## <none>                                  26.693 -27901
## - Macromoths                1    0.0297 26.722 -27897
## - Bees                      1    0.0894 26.782 -27885
## - Butterflies               1    0.4412 27.134 -27816
## - Grasshoppers_._Crickets   1    0.4566 27.149 -27813
## - Ladybirds                 1    1.3177 28.010 -27648
## - Vascular_plants           1    3.7348 30.428 -27211
## - Hoverflies                1    6.9221 33.615 -26686
```

Macromoths, Bees, Butterflies, Grasshoppers_ Crickets, Ladybirds, Vascular Plants, and Hoverflies are arranged from least to most significant in terms of their contribution to model fit.

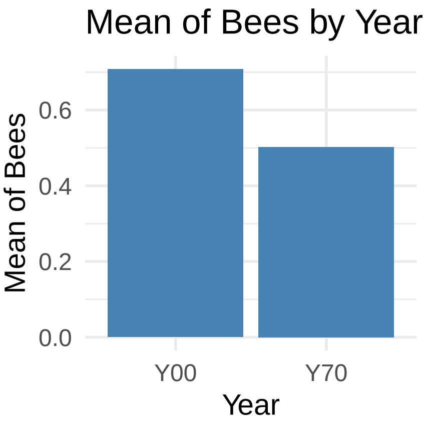
Open analysis

Bees vs Period

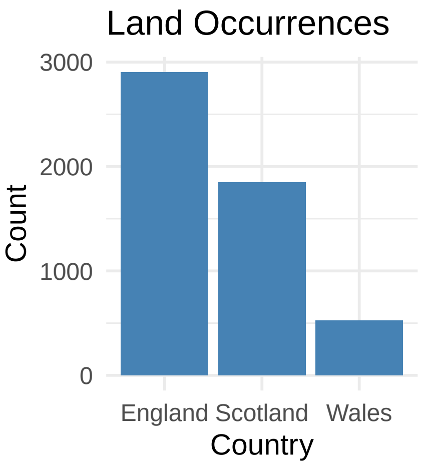
After running a separate t-test to compare the mean bee populations of 'y00' and 'y70', we found evidence of a significant shift in the bee population.

```
##
##  Two Sample t-test
##
## data:  Main_Proj_data$Bees by Main_Proj_data$period
## t = 25.465, df = 5278, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Y00 and group Y70 is not equal to 0
## 95 percent confidence interval:
##  0.1896400 0.2212737
## sample estimates:
## mean in group Y00 mean in group Y70
##         0.7077523         0.5022954
```

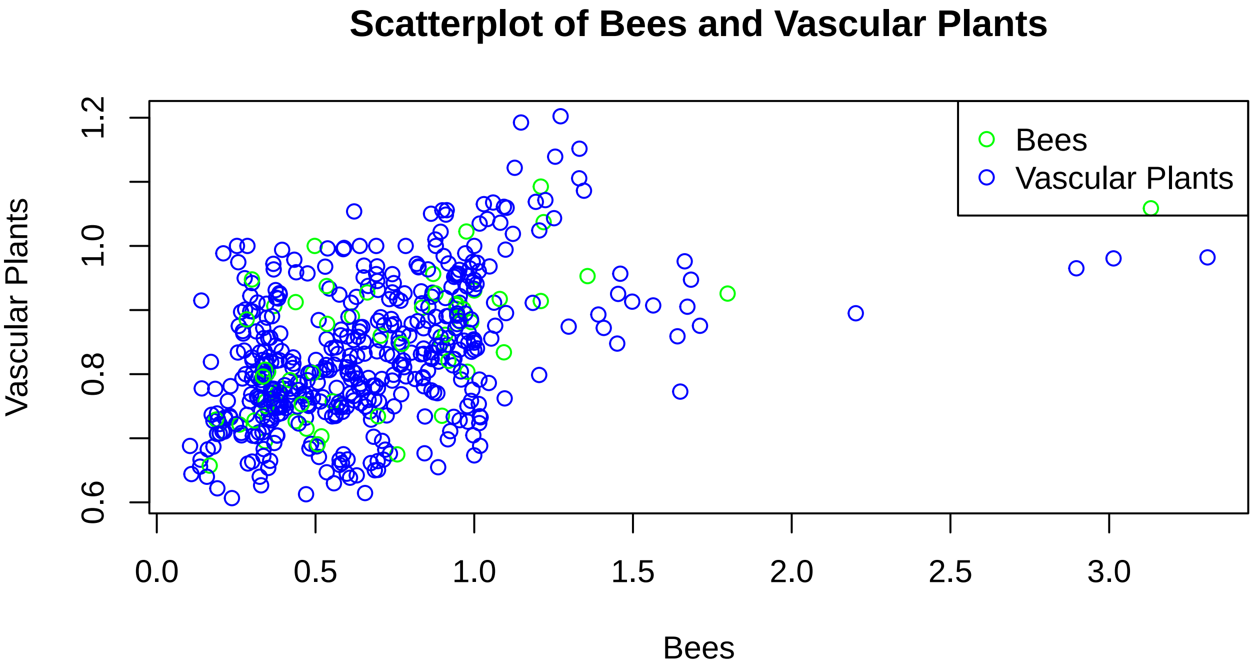
There was a statistically significant change in the bee population between the two time periods, as measured by a t-value of 25.465 and a p-value of less than 2.2e-16. The 95% confidence interval of (0.1896400, 0. 2212737) further supports the idea that the real mean difference is larger than zero. In the 'y00' period, the average number of bees was calculated to be 0.7077523, which is much greater than the 'y70' period average of 0.5022954. This points to a drastic decline in bee numbers between the two periods.



To ensure the long-term viability of the bee population, this data is essential for environmental and conservationists to take the appropriate action.



When we performed analysis furthermore on the Dominant land class, we grouped all the occurrences and found out that England had the most number of occurrences with about 3000, however Wales had the least with only around 500. This maybe because of the area of land available which we all know that the area of England was more compared to that of Scotland and Wales.



We compared the Bees and Vascular Plants. The findings demonstrated that bees and Vascular Plants were related very well. The analysis showed strong evidence that wherever we found the existence of Vascular plants the existence of bees were precise too. This further tells us that we can find most of the bees population where we have vascular plants. (Figure 7) Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

taxi_group	Mean	Median	SD	Skewness
Bees	0.68	0.63	0.39	2.18
Butterflies	0.89	0.89	0.07	-0.29
Hoverflies	0.75	0.75	0.13	0.04
Ladybirds	0.74	0.75	0.21	-0.12
Macromoths	0.94	0.94	0.08	0
Grasshoppers_._Crickets	0.59	0.56	0.25	0.33

taxi_group	Mean	Median	SD	Skewness
Vascular_plants	0.83	0.82	0.11	0.43

This table presents quantitative measures, including mean, median, and standard deviation, that have been calculated for the BD7 Taxonomic group. These measures provide insight into the ecological status of each species in the group. Additionally, ANOVA or regression analyses have been performed to examine any variances across the groups.

Below are the observation: Macromoths exhibit a more symmetrical distribution with a skewness value close to zero compared to other groups. Butterflies and Ladybirds have a negative mean distribution, suggesting that the distribution of species within these groups is more biased towards the negative end of the mean.