MA334_2309756

Univariate analysis and basic Statistics:

taxi_group	min	1st Q	mean	median	3rd Q	Max	Winsorized Mean
Bird	0.371724	0.872026	0.92	0.91919	0.967185	1.171986	0.92
Butterflies	0.436916	0.78253	0.88	0.893365	0.969194	1.394366	0.88
Carabids	0.077356	0.562407	0.68	0.690413	0.795224	1.199766	0.68
Grasshoppers_._ Crickets	0.119617	0.520884	0.66	0.651252	0.811927	1.59375	0.66
Vascular_plants	0.538308	0.712139	0.78	0.779522	0.850156	1	0.78

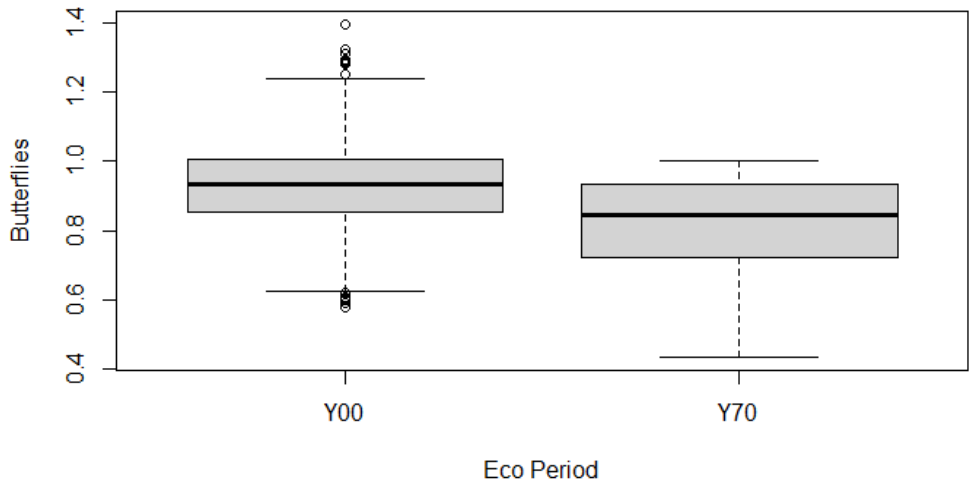
The above table consist of the five species of richness which has been explored and analysed for the location England. Key stats Like Mean, Median, 1st Quartile, 3rd Quartile, Winsorized Mean has been calculated. Higher mean values of **Birds** and **Bees** indicates a significant association between the species distribution throughout the location. Alternatively, **Carabids** and **Grasshoppers_._Crickets** has lower mean values indicating a slightly lower association between the species distribution through the location.

There is no significant difference between the mean and winsorized mean. This signifies that replacing the first and last 20% of the values is not affecting the true mean.

Correlation Table	Bird	Butter flies	Carabids	Grasshoppes _._Crickets	Vascular_ plants
Bird	1	0.23	0.26	0.23	0.25
Butterflies	0.23	1	-0.05	0.36	-0.07
Carabids	0.26	-0.05	1	0.39	0.45
Grasshoppers_._ Crickets	0.23	0.36	0.39	1	0.38
Vascular_plants	0.25	-0.07	0.45	0.38	1

The above correlation table shows the weak and moderate correlation between the Richness Species. The highest correlation being +0.45 between the **Vascular Plants** and **Carabids**.

From the boxplot, the specific richness species is normally distributed across the locations in the later time period Y00.But on Y70 period there is a positive skew noticed. The outliers are noticed in the later period and not in earlier period. There variability has reduced in the in the later period compared to the initial period.

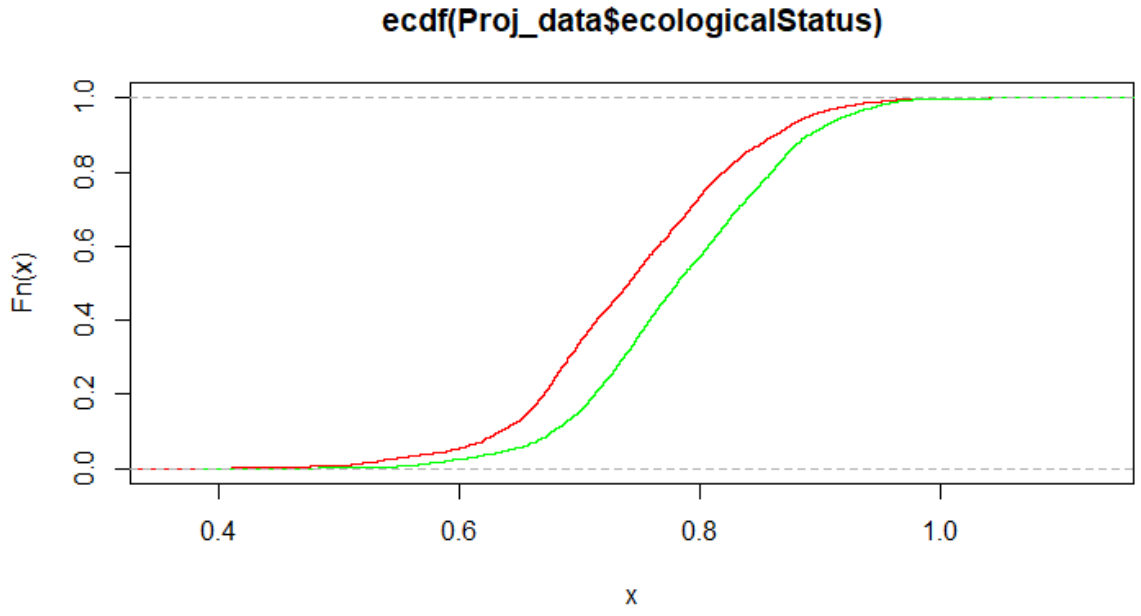


Hypothesis testing's:

Kolmogorov-Smirnov Test:

The Kolmogorov-Smirnov test resulted in D value of 0.19972 through which obtained p-value (2.2e-16) is less than 0.05, indicating that the Null hypothesis should not be supported. This suggests strong evidence that the distributions are different. In similar terms, there's clear support for the idea that the distributions are not the same.

The graph below shows the cumulative distribution functions (CDFs) for "**Ecological Status**"(indicated by the red curve) and **"eco_status_5"**(highlighted in green) show a exact view of both CDFs and the difference in it. From the calculated analysis, the two datasets does not significantly overlap and both the CDFs are different.



The Kolmogorov-Smirnov does not depend on assumptions about the distribution of the data. This characteristic makes it more advantageous when comparing distributions that change from standard or normal patterns.

One Sample T-Test:

From the T-test conducted data variable "**BD5 change**", p-value less than 0.05 indicates enough support to not consider null hypothesis. This implies that there is sufficient evidence to conclude that the mean is not equal to zero.

The T value obtained (-2.8181) indicates the difference between the observed sample mean and the hypothesis mean value (0), and the difference and uncertainty in the sample mean. The larger t-value indicates a larger difference between the sample mean and the hypothesis mean, indicates the stronger evidence against the null hypothesis.

The t-test is adaptable and widely used statistical method, especially in circumstances comparing statistics or conducting hypothesis tests, dealing with small sample sizes, or when the population SD (Standard Deviation) are unknown.

Contingency table/comparing categorical variables:

Contingency table

	BD5_Down	BD5_Up	Total
BD11_Down	496	341	837
BD11_Up	216	399	615
Total	712	740	1452

Independent table

	BD5_Down	BD5_Up	Total
BD11_Down	410	427	837
BD11_Up	302	313	615
Total	712	740	1452

Likelihood and Odds ratio:

The log likelihood ratio(G-test) taken for Independence on the Contingency table has given a G-value of 83.617 along with p-value lower than 0.05, indicating that there is strong evidence that the null hypothesis should not be supported. This indicates that the variables in the table are not independent. In simple terms, if a value in the table increases/decreases the other corresponding variable also increases/decreases as they are dependent variables.

The Odds Ratio 2.686869 suggest an association between the two events (that event being up or down). The value indicates that the odds of the event being “up” is 2.686869 times higher than the odds of event being “down”. The chances of the event being “up” is very high as the odds ratio is greater than 1.

Sensitivity, Specificity and Youden Index:

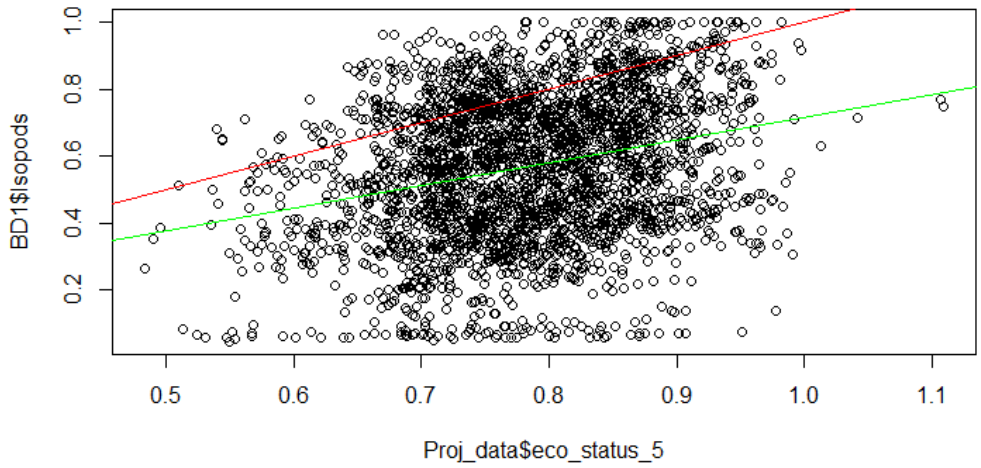
The Sensitivity is carried out to show its ability to detect the true positive outcome, which is approximately 69.66% (value obtained from the test is 0.6966292). Similarly, Specificity can

be used to detect the true negative outcome, which shows approximately 53.92% (value obtained from the test is 0.5391892). From the above test, it's clear that they display a reasonable ability to detect true positive outcome, but its ability to reject the true negative values is comparatively low. The value obtained through Youden Index is 0.236 suggest the moderate value obtained for the diagnostic test indicating the balanced performance.

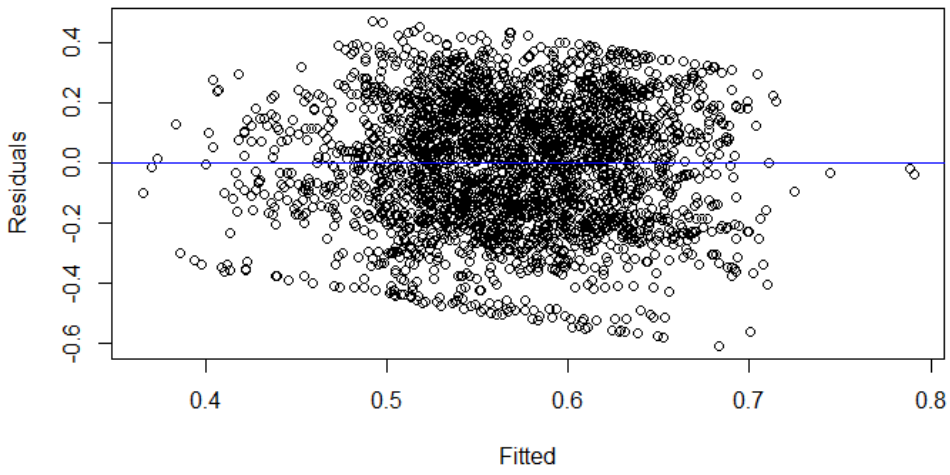
Simple Linear Regression:

Taking linear Model relating Isopods as a response variable to the predictor variable **eco_status_5** has given many insights. A single value increase in the **eco_status_5** is tending to be 0.67886 increase in the Isopods, forming a slope on graphical view of the linear model.

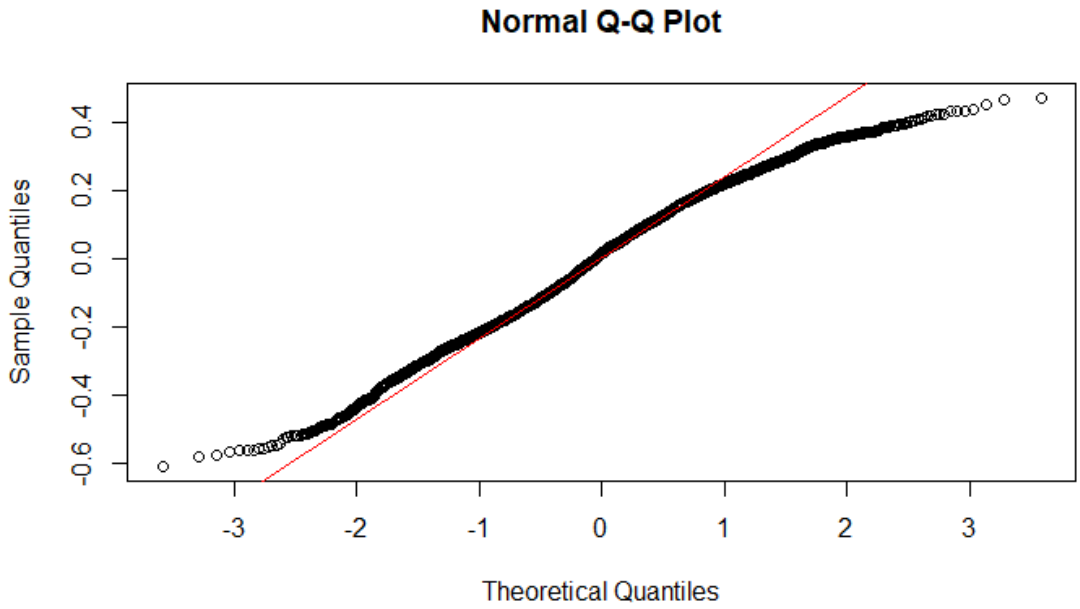
The p-value obtained from the linear model is smaller than significant level, indicating a strong relationship between the Isopods and eco_status_5. This implies that if any variation in the **eco_status_5** said to be noticed will also be noticed in the Isopods. The Adjusted R-Square value is 0.07 indicate very poor model fitting.



The plot-1 represents clear deviation between zero line (highlighted in Red) and fitted line (highlighted in Green). Data points are evenly(approx.) scattered between zero line and fitted line. And it's seen that many data points are far scattered from the zero line. There is a relation as mentioned earlier that if single value increases in the eco_status_5, Isopods increases by 0.67886 times. This can be noticed in the Plot1.



Plot-2 represents fitted vs residuals. The data points are densely plotted away from the zero line, indicating higher error in model.



Plot-3(Q-Q Plot) shows poor overlap between data points and the zero line. There is significant number of outliers in the dataset.

Multiple Linear Regression:

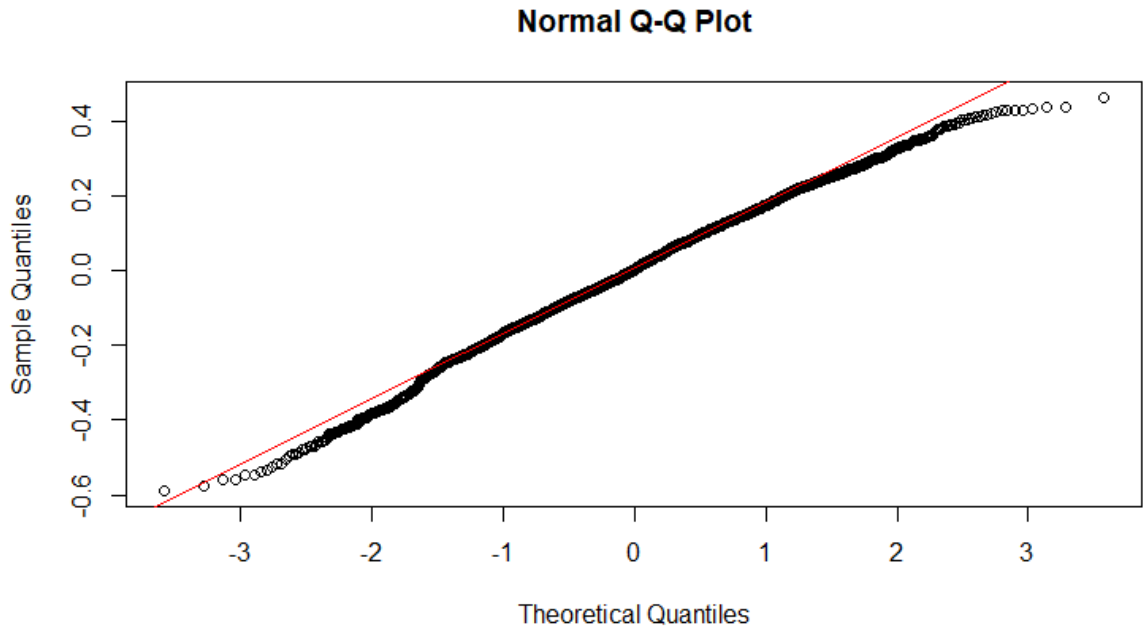
	Df	AIC
lmMod	7	-1871.01
lmMod_reduced	6	-1872.71

AIC value for the initial MLR is -1871.012 and the corelation value is 0.5812951. Initially multiple linear regression model has Isopods as a response variable while the five selected taxonomic groups are predictor variable. The initial MLR for the five taxonomic groups clearly showed that the **butterflies** have higher p-value.

Reduced Model removing **butterflies** from the five taxonomic group has better AIC model value. Hence the reduced model performs much better than the initial MLR Model as the AIC is low, and P Value is less than 0.05.

	Df	AIC
lmMod	7	-1871.01
lmMod_reduced	6	-1872.71
lmMod_interaction	8	-1874.95

On an iterative basis, the interaction linear model interacting between Vascular plant and **Grasshopperss_._Crickets** shows more acceptable AIC value of -1874.946. While interacting every taxonomic group with each other, the above groups have the lowest AIC model values and performs significantly better compared to the initial MLR model and Reduced model. The corelation value has slightly increased compared to the initial MLR and Reduced MLR Model. The corelation values obtained is 0.5824563. Also, the Adjusted R Square value obtained from the interaction model is 0.3379, slightly better than the other models.



From the above plot, the data points are normally distribution. Hence, these data suggest that the Interaction Model performs well and can be a better fit compared to the initial and reduced MLR models.

Training and Test Set:

The Linear Regression Model shows high significance of **Birds**, **Butterflies**, **Carabids** and **Grasshoppers_._Crickets** with respect to the **Isopods**. The Model has an adjusted r square value of 0.2153 with F Score of 145.8.

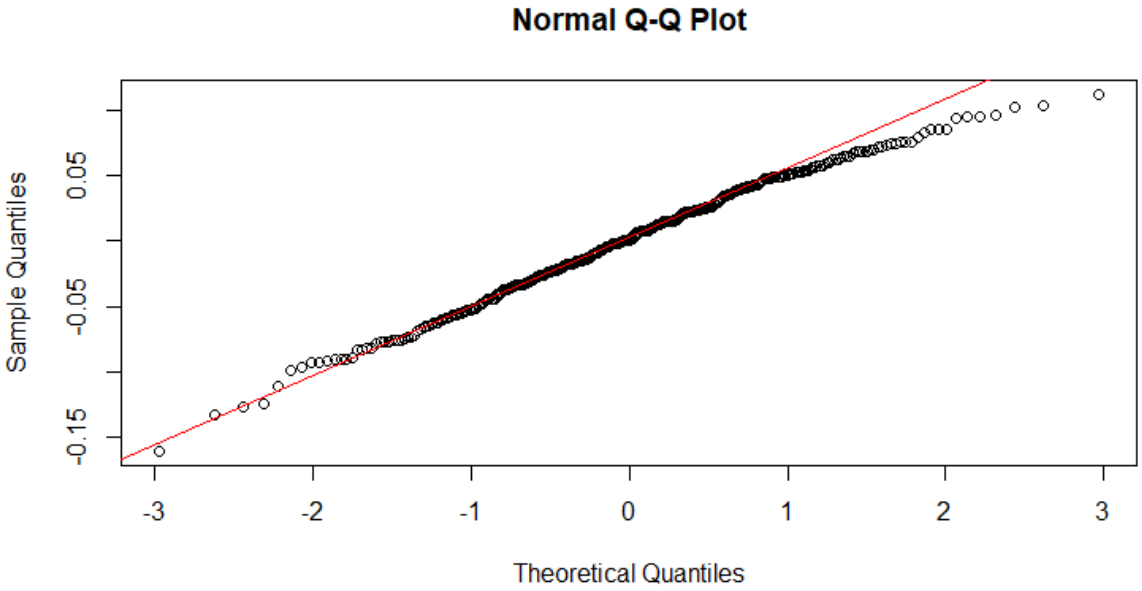
The initial training dataset had an MSE of 0.02608887. The test dataset run on the initial trained model yielded a MSE of 0.0676, which is higher than the MSE of training dataset.

The initial trained dataset always has a lower MSE value as the data is well known and well trained by the model. But the test dataset is a new dataset which is not familiar for the model to run and provides output with more error than the trained data set. This leads to a higher MSE value for the test data set.

Open Analysis:

In Open analysis, A linear regression model has been generated for the **eco_status_5** (that's five taxonomic groups) as response and Dominant Land Class as predictor variable. This analysis has been done for the selected five taxonomic groups for specific Location in England to precise analysis the species richness. The chosen location is centre parts and west of the England for the specific period Y00.

The Open analysis has a F-value of 39.01 along with the P-value 2.2e-16 which is less than 0.05. This states that the variable has strong relationship with the eco_status_5. The adjusted R square value obtained is 0.554. Changes in dependent variable can be explained 50% (approx.) by the independent variable. The correlation value for the model is 0.754351. From the Linear model, the dominant land class 15e, 17e and 3e are said to have higher p-value. The valleys, undulating plain which are slight up and down like a wave and slope hills makes an unconditional place to live for species richness like insects, birds, and plants. The other dominant land classes are highly significant.



From the above plot, most of the data points lies on the line. There is slight spread in the data points closer to the line. The data points are normal distributed. Hence the Open Analysis has a significant relationship between the taxonomic groups and dominant class for the later period Y00.