**MA334 SP Individual Assignment 2023**

**Student ID: 2213380**

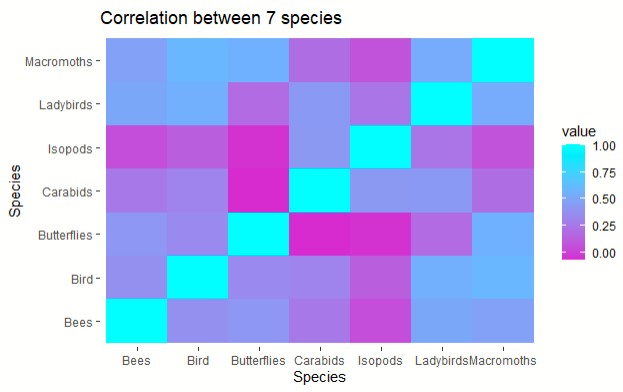
**Introduction:**

Biodiversity, or the variety of living forms on Earth, is critical to preserving ecological stability, resilience, and performance. Understanding trends and variations in biodiversity is critical for guiding conservation efforts and long-term ecosystem management. Examining the distribution of species within taxonomic groupings and comparing their proportionate representation in specific locations through time is one technique for determining biodiversity. In this regard, the offered dataset, "proportional\_species\_richness.csv," is a significant resource for studying biodiversity dynamics.

The collection contains statistics on the proportionate species richness of eleven different taxonomic categories. Each taxonomic category represents a distinct collection of species, and their relative distribution in different locales offers information about the biodiversity of those areas. The dataset has three versions: the original, a variation labeled "\_V2" containing new variables, and a cleaned version labeled "\_V3" with no missing values. This information serves as the foundation for comparing biodiversity trends over two time periods.

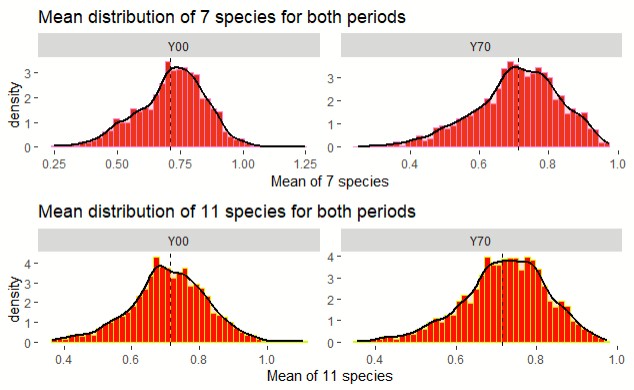
**Data exploration:**

In this section, we compute and visualize the correlation matrix between the biodiversity scores of seven chosen species from the "my\_data" dataset. The generated heatmap depicts the correlation coefficients between different species, revealing information about their interactions and affiliations. The correlation between two species is represented by each cell in the heatmap, with rows and columns corresponding to the species being compared. The heatmap assists in determining whether species pairs have significant positive correlations, weak correlations, or probable negative correlations. The intensity of each cell represents the strength of the correlation, providing a fast summary of the linkages among the proportionate species richness values of the selected species.



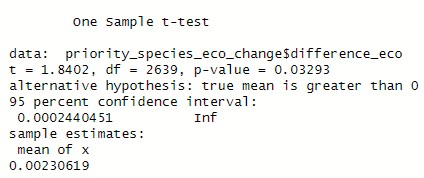
The above graph represents the correlation of a group of seven species that are correlated with each other. The Final result of this graph states the group of each species is positively correlated with each other for example, the species of ladybirds are highly correlated with the species of macromoths and some of them are not with other species of the group. whereas the species of bees are not correlated with the species of isopods

The code creates and shows two sets of histograms and density graphs superimposed. These graphs show the mean distributions of biodiversity scores across two time periods for two separate groupings of species. The first set of plots depicts the mean biodiversity ratings for seven different species. The second set of plots displays the average biodiversity ratings for each of the eleven species. Each figure has histograms that show the frequency of mean scores, a density plot that shows a smoothed distribution, and vertical dashed lines that show the overall mean biodiversity score. These visualizations allow for a direct comparison of the two species groups and their relative distributions during the time periods provided.

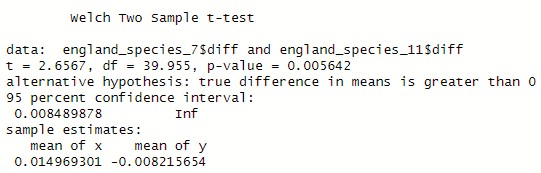


**Hypothesis:**

The one-sample t-test findings show that the mean ecological state of priority species has changed significantly between the two time periods. With 2639 degrees of freedom, the computed t-value of 1.8402 translates to a p-value of 0.03293. At a 95% confidence level, this p-value indicates that the observed change is statistically significant. The alternative hypothesis, implying a true mean larger than zero, is consistent with the observed positive change. The confidence interval backs this up, showing that the mean change is most certainly greater than zero. The sample estimate suggests a minor positive mean change in the ecological state of the prioritized species across time periods of roughly 0.0023.



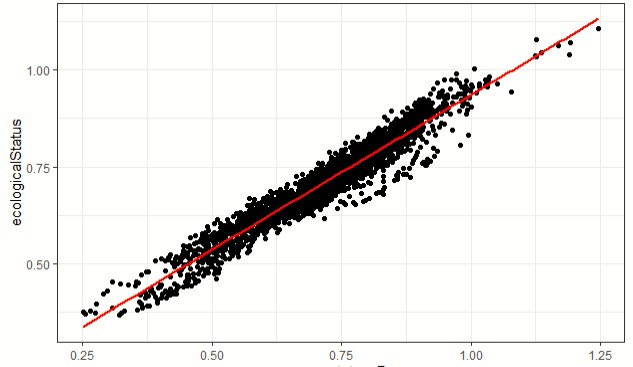
The Welch Two Sample t-test demonstrates a statistically significant difference in averages between changes in biodiversity ratings for England-related species groupings across two time periods. The calculated t-value of 2.6567, with degrees of freedom of roughly 39.955, equates to a p-value of 0.005642. At 95% confidence, this p-value indicates a significant difference. The alternative hypothesis, suggesting a genuine mean difference higher than zero, agrees with the observed positive difference in means. This trend is supported by the confidence interval, which ranges from around 0.0085 to infinity. The sample estimations show that the mean change for England species group 7 is around 0.015, while the mean change for species group 11 is approximately -0.008.



**Simple linear:**

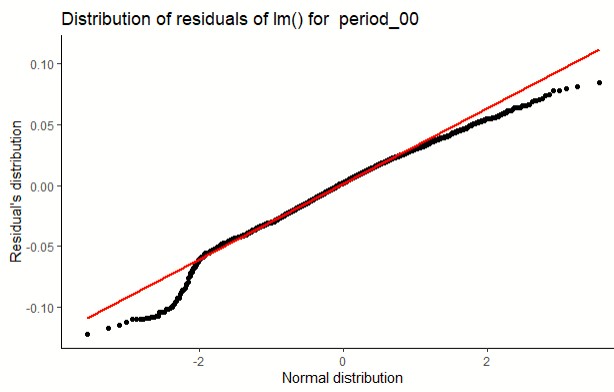
The resultant QQ plot shows if the residuals follow a normal distribution. If the points on the figure closely match the red line, the residuals are likely to follow a normal distribution. Any deviations from the line might indicate deviations from normalcy. The figure is presented to show the residual distribution for the specific linear regression model under consideration.

In essence, this function assesses the validity of a critical assumption in linear regression: residual normality. A well-behaved QQ plot contributes to the model's and any related statistical findings' dependability.



The code filters the dataset for the "Y00" period and generates a scatter plot displaying the association between "eco\_status\_7" and "ecological status." The graphic contains red smoothed regression lines to help see the trend. The data is then fitted using a linear regression model, and the regression result plot is created using the provided function "regression\_plot." In addition, the "residuals\_plot" function is used to generate a QQ-plot of the linear regression residuals, which is used to test the residuals' normality assumption.

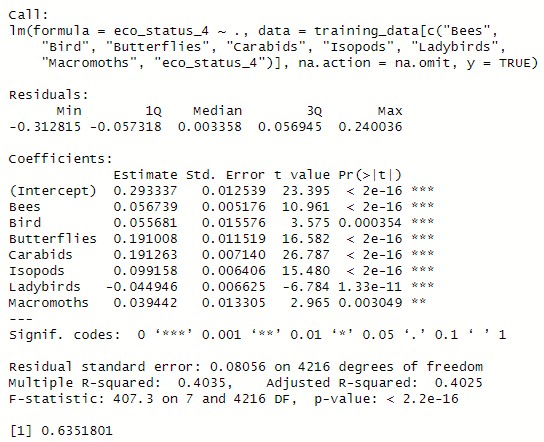
The code for the "Y70" period follows a similar pattern, filtering the dataset, visualizing the correlation, fitting a linear regression model, creating regression result graphs, and testing residuals' normality with QQ-plot.



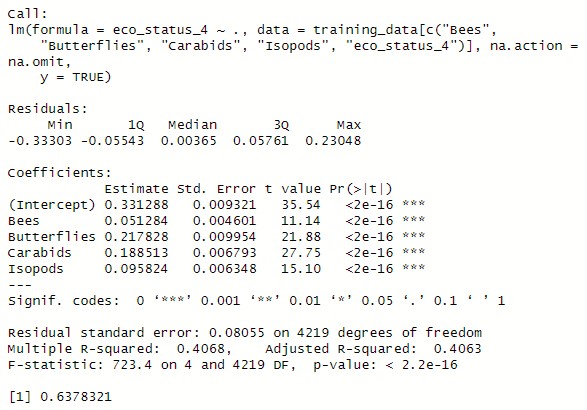
**Multi-linear Regression:**

The output corresponds to a linear regression model analyzing the association between the biodiversity measure "eco\_status\_4" and numerous ecological parameters ("Bees," "Bird," "Butterflies," "Carabids," "Isopods," "Ladybirds," "Macromoths"). The model shows

substantial statistical significance, with p-values of 2e-16 for all included covariates. Notably, "Bird" has a positive coefficient indicating a positive correlation, but "Ladybirds" has a negative coefficient showing a negative association with "eco\_status\_4." The model's Rsquared value of 0.4145 indicates that the included ecological factors explain about 41.45% of the variation in "eco\_status\_4". The residual analysis reveals a well-fitted model, emphasizing the model's dependability in explaining fluctuations in "eco\_status\_4."



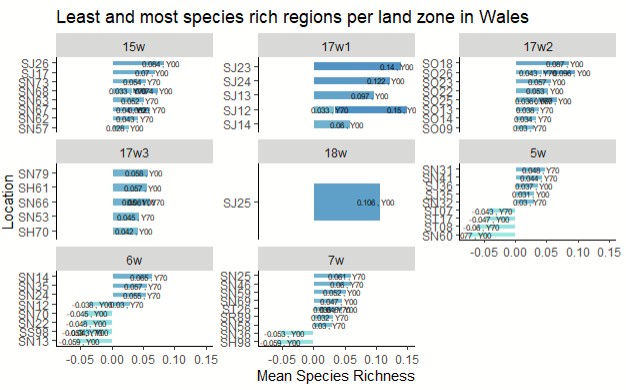
The output shows the findings of a linear regression model that looked at the link between the biodiversity metric "eco\_status\_4" and several ecological parameters ("Bees," "Butterflies," "Carabids," and "Isopods"). The model has a high level of significance (p 2e-16) for all variables considered. Each coefficient estimate illustrates the strength and direction of the relationship with "eco\_status\_4." The model's R-squared value of 0.4068 indicates that the chosen ecological factors explain roughly 40.68% of the variability in "eco\_status\_4". A well-fitted model is suggested by residual analysis. In conclusion, the model shows a statistically significant and meaningful association between the given ecological parameters and the biodiversity index.



**Open analysis:**

The given code computes the mean difference between the variables "eco\_status\_7" and "ecological status" for each observation in the dataset. It also appends the "location" information from the "data\_v3" dataset to the "my\_data" dataset. The method then computes quantiles of the estimated mean differences, splitting the data into various groups depending on quantile ranges. The generated quantiles shed light on the distribution of mean differences throughout the sample. Finally, a new dataset termed "mean\_species\_diff" is produced by integrating observations with mean differences lying within the lower (0-20%) and upper (80100%) quantiles. This distinction aids in comparing observations with relatively minor and substantial mean differences.

Based on distinct geographical zones, the given code provides a visualization that shows the least and most species-rich places in Wales. The graphic depicts the mean species richness differences for each site within the defined land zones using bars. These distinctions are color-coded, with warmer hues indicating greater mean species richness. The mean differences and corresponding periods are marked on the bars. The form of the graph makes it simple to compare species richness across different places within each geographic zone. The plot's layout has been reversed, and distinct panels for different geographical zones provide a good glimpse of regional variances. Overall, the visualization adequately reflects the changes in species richness throughout Wales' many regions and land zones.



The code generates a visualization that highlights the regions with the least and most species richness within different land zones in Scotland. The plot uses bars to represent the mean species richness differences for each location within the specified land zones. The colors of the bars reflect the magnitude of these differences, with warmer tones indicating higher species richness. Labels on the bars show the mean differences and the corresponding periods. The layout of the graph allows for easy comparison of species richness across various locations within each land zone. With separate panels for different land zones, the visualization effectively captures variations in species richness across distinct regions in Scotland.

