# Data Analysis on Insects Data

Musaddik Maulavi

2024-01-12

## Introduction

The statistical analysis and exploratory data analysis is being performed on the dataset having species richness and the dominant land class. The method first generates lists all the species, and then the allocation of the selected five taxonomic group (Butterflies, Carabids, Hoverflies, Ladybirds, Grasshoppers_._Crickets). These five taxonomic group is termed as BD5 in the code. The program then computes minimum, maximum,Q1, Q2, Mean, Median, winsorized mean to examine the data further. The code then creates the correlation matrix between all pairs of variables in BD5 and it is represented in a tabular form. Then boxplot is created for variable 'Hoverflies'. Further, the code computes Hypothesis test with two tests 'T- Test' and 'KS – Test', generating a p-value as an output. Then, the contingency test is performed with BD5_up against BD11_up. With the help of contingency test, the code calculates Odds ratio, Sensitivity, Specificity and Youden's index. The script then executes linear regression and multiple linear regression. The key objective of code is to clarify the links between the diversity of species and land class, to find trends or patterns, and to investigate possible reasons for variations in biodiversity indexes throughout time.

## Summary of BD5 (5 Variables)

```
table <- data.frame()
for(i in c(1:5)){
  table <- rbind(table,
                 c(names(BD5)[i],
                   round(min(BD5[,i],na.rm = TRUE),digits = 2),
                   round(quantile(BD5[,i], 0.25),digits = 2),
                   round(median(BD5[,i],na.rm = TRUE),digits = 2),
                   round(mean(BD5[,i],na.rm = TRUE),digits = 2),
                   round(quantile(BD5[,i], 0.75),digits = 2),
                   round(max(BD5[,i],na.rm = TRUE),digits = 2)
                 ))
}

# Calculating Winsorised Mean
winsorised_mean <- round(sapply(BD5 , function(x) mean(Winsorize(x , 0.2))), digits = 2)

winsorised_mean <- as.data.frame(winsorised_mean)

winValue <- winsorised_mean %>% pull(winsorised_mean)

# Binding the winsorised value to summary table created above
main <- cbind(table,as.data.frame(winValue))

colnames(main) <- c("Names","Minimum","Q1","Median","Mean","Q2","Max","Winsorized")

main
```

```
##                     Names Minimum   Q1 Median Mean   Q2  Max Winsorized
## 1             Butterflies    0.32 0.79   0.89 0.87 0.97 1.39       0.87
## 2                Carabids    0.01 0.48   0.64 0.61 0.76  1.2       0.61
## 3              Hoverflies    0.12 0.57    0.7 0.68 0.81 1.15       0.68
## 4               Ladybirds    0.06 0.45   0.64 0.61  0.8 1.84       0.61
## 5 Grasshoppers_._Crickets    0.07 0.49   0.62 0.63 0.79 1.59       0.63
```

The above table shows summary data and Winsorized values for five ecological groups: butterflies, carabids, hoverflies, ladybirds, and grasshoppers & crickets. Each group's statistics have the minimum value, first quartile (Q1), median, mean, third quartile (Q2), and maximum values. Winsorized values are a type of outlier treatment. Particularly, Hoverflies have the highest median and maximum values among the categories, whereas Carabids have the lowest minimum value. Ladybirds had the greatest maximum value, which might indicate increased variability. These summary statistics shed light on the central tendency and variability within each ecological category, facilitating ecological and statistical research.

## Correlations Between Variables In BD5

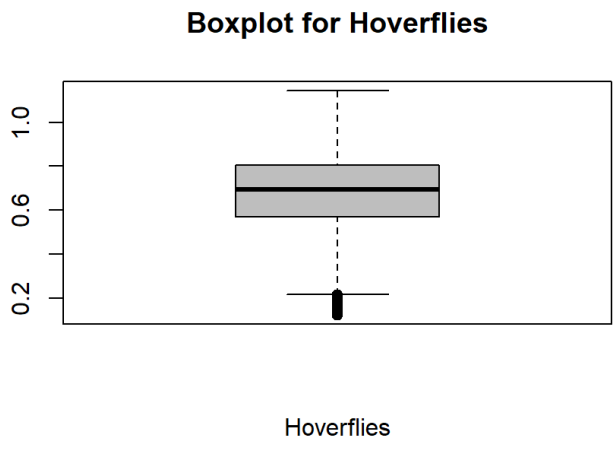
```
# Creating a correlation matrix for BD5 variable
correlation_matrix <- cor(BD5)
colnames(correlation_matrix)<-c("Butterflies","Carabids","Hoverflies","Ladybirds","GrassHop&Crics")
rownames(correlation_matrix) <-c("Butterflies","Carabids","Hoverflies","Ladybirds","GrassHop&Crics")

correlation_matrix
```

```
##                Butterflies    Carabids Hoverflies Ladybirds GrassHop&Crics
## Butterflies     1.00000000 -0.07224957  0.2797315 0.1851943      0.2425145
## Carabids       -0.07224957  1.00000000  0.4966727 0.4302754      0.4532495
## Hoverflies      0.27973153  0.49667268  1.0000000 0.3989119      0.4223757
## Ladybirds       0.18519428  0.43027541  0.3989119 1.0000000      0.3092311
## GrassHop&Crics  0.24251447  0.45324951  0.4223757 0.3092311      1.0000000
```

The diagonal values (from top left to bottom right) show each variable's perfect correlation with itself, which is always 1.00. Hoverflies exhibit positive correlations with a range of species, including butterflies, Carabids, ladybirds, and grasshoppers/crickets, suggesting an ecological link or shared environmental preferences.The strong correlation with Carabids, in particular, may indicate a deeper ecological link or shared habitat preferences between hoverflies and carabids. Carabids have a moderately positive association with other species and a strong positive correlation with Hoverflies Butterflies shows notable positive correlations with 'Hoverflies', 'Ladybirds', and 'Grasshoppers_._Crickets', yet just a small negative relationship with 'Carabids'. It can be seen that there is a fairly positive relationship between Ladybirds, Grasshoppers_._Crickets, and Hoverflies

## Box-plot for variable 'Hoverflies' from BD5

```
#summary(BD5$Hoverflies)
boxplot(BD5$Hoverflies,
        main = "Boxplot for Hoverflies",
        xlab = "Hoverflies",
        ylab = "",
        col = "grey",  # Change the color if desired
        border = "black")
```




The above boxplot of Hoverflies depicts that the minimum value of Hoverflies is approximately 0.12 and maximum value is around 1.14. It also shows that the 1st quartile is around 0.57 and 3rd quartile is around 0.81 which indirectly depicts that median lies in 0.69 approximately.

## Hypothesis Test - T-Test

```
t_test <- t.test(BD5$Ladybirds,BD5$Grasshoppers_._Crickets)

# Printing the p-value from the t-test
#print(t_test)
print(t_test$p.value)
```

```
## [1] 0.001451213
```

The p-value which is obtained from t-test (0.001451) is less than the significance level of 0.05. It indicates strong evidence against the null hypothesis, suggesting a significant difference in means between 'Ladybirds' and 'Grasshoppers_Crickets'. It shows that it has sufficient evidence to reject the null hypothesis and declare that there is a statistically significant difference in the means of 'Ladybirds' and 'Grasshoppers_Crickets' based on this Welch Two Sample t-test.

## Hypothesis Test - KS-Test

```
ks_test_result <- ks.test(BD5$Ladybirds, BD5$Grasshoppers_._Crickets)
```

```
## Warning in ks.test.default(BD5$Ladybirds, BD5$Grasshoppers_._Crickets): p-value
## will be approximate in the presence of ties
```

```
# Print the KS test result
#print(ks_test_result)
print(ks_test_result$p.value)
```

```
## [1] 1.110223e-16
```

With such a small p-value (1.110223e-16), there's strong evidence to reject the null hypothesis. It implies that whatever impact or link the statistical test was looking into is highly unlikely to be due to chance or randomness. Thus, we can conclude that the distributions of 'Ladybirds' and 'Grasshoppers_Crickets' significantly differ based on the Asymptotic Two-Sample Kolmogorov-Smirnov test.

## Contingency Test

```
# The mean of values of BD5 is stored in the vector form using rowMeans
BD5_mean_selected <- rowMeans(proj_data[,my_variables])
# Selecting my 5 variables and adding another column of BD5 mean selected
Proj_data <- proj_data%>%select("Location",all_of(my_variables),
                                      "Easting","Northing","dominantLandClass",
                                      "ecologicalStatus","period")%>%mutate(eco_status_5=BD5_mean_selected)

# For BD5, Mutating another column BD5 change which will be the dufference between two periods
Proj_data_split <- Proj_data%>%select(Location,period,eco_status_5)%>%
  pivot_wider(names_from =period,values_from=eco_status_5)%>%
  mutate(BD5_change=Y00-Y70)
# View(Proj_data_split)


# Extracting BD5_change and storing into BD5_change variable
BD5_change <- Proj_data_split%>%pull(BD5_change)


#########################

# Same for BD11,  Mutating another column BD5 change which will be the difference between two periods
proj_data_split <- Proj_data%>%select(Location,period,ecologicalStatus)%>%
  pivot_wider(names_from =period,values_from=ecologicalStatus)%>%
  mutate(BD11_change=Y00-Y70)
#hist(proj_data_split$BD11_change)  # the distribution of the BD5 change

# Extracting BD11_change and storing into BD11_change variable
BD11_change <- proj_data_split%>%pull(BD11_change)


Eco_change_BD11 <- proj_data_split%>%select(Location,BD11_change)
Eco_change_BD5 <- Proj_data_split%>%select(Location,BD5_change)
# Combining both BD11 and BD5
Both_eco_change <- inner_join(Eco_change_BD11,Eco_change_BD5,by="Location")

# Mutating BD11 up and BD5 up,if the respective change value is positive then the value is set to Increase otherwise Decreas
e.
Both_eco_change <- Both_eco_change%>%
  mutate(BD11up=ifelse(Both_eco_change$BD11_change>0,"Increase","Decrease"))%>%
  mutate(BD5up=ifelse(Both_eco_change$BD5_change>0,"Increase","Decrease"))

BD5up<-table(Both_eco_change$BD5up)
BD11up<-table(Both_eco_change$BD11up)


Table_up_down <- table(Both_eco_change$BD11up,Both_eco_change$BD5up) # contingency table for interpretation
# Changing row names and column names
colnames(Table_up_down) <- c("Decrease","Increase");rownames(Table_up_down) <- c("Decrease","Increase")
Table_up_down
```

```
##
##            Decrease Increase
##   Decrease     1332      306
##   Increase      355      647
```

```
BD5up # Table for BD5 up
```

```
##
## Decrease Increase
##     1687      953
```

```
BD11up # Table for BD11 up
```

```
##
## Decrease Increase
##     1638     1002
```

The above table shows the observed counts, where BD11 decreased 1332 times, followed by BD5, which decreased 1332 times and increased 306 times. BD11 increased 355 times, with BD5 decreasing 355 times and increasing 647 times. The increase and decrease values are calculated by finding the differences of BD5 and BD11 values between the two periods i.e Y70 and Y00. This data depicts the link between BD11 and BD5 alterations, specifically how frequently they increase or decrease together. It is critical to assess these counts in relation to the expected counts in order to determine if BD11 and BD5 alterations are independent or associated.

## Likelihood Ratio Test

```
# Create matrices for the contingency tables (One for BD5 against BD11 and other for Independent model)

independent_table <-rbind(table(Both_eco_change$BD5up),table(Both_eco_change$BD11up))  # Assuming only BD11up is considered in the independent model
rownames(independent_table) <- c("BD5up","BD11up")

table1 <- matrix(c(1687, 953, 1638, 1002), nrow = 2, byrow = TRUE)# Independent Contingency table
table2 <- matrix(c(1332, 306, 355, 647), nrow = 2, byrow = TRUE)# BD5up against BD11up


GTest(table1)
```

```
##
##  Log likelihood ratio (G-test) test of independence without correction
##
## data:  table1
## G = 1.9504, X-squared df = 1, p-value = 0.1625
```

```
GTest(table2)
```

```
##
##  Log likelihood ratio (G-test) test of independence without correction
##
## data:  table2
## G = 572.68, X-squared df = 1, p-value < 2.2e-16
```

```
p_value1 <- GTest(table1)$p.value
p_value2 <- GTest(table2)$p.value


# Compare p-values with the significance level (alpha = 0.05 for 5% confidence level)
alpha <- 0.05
comparison1 <- ifelse(p_value1 < alpha, "Reject Null Hypothesis", "Fail to Reject Null Hypothesis")
comparison2 <- ifelse(p_value2 < alpha, "Reject Null Hypothesis", "Fail to Reject Null Hypothesis")

cat("Comparison for Independent model Table having p value - ",p_value1, comparison1, "\n")
```

```
## Comparison for Independent model Table having p value -  0.1625445 Fail to Reject Null Hypothesis
```

```
cat("Comparison for Contingency Table BD5up against BD11up: having p value ",p_value2, comparison2, "\n")
```

```
## Comparison for Contingency Table BD5up against BD11up: having p value  0 Reject Null Hypothesis
```

For table 1, the test generated a statistic (G) of 1.9504 with 1 degree of freedom, providing a p-value of 0.1625. With a standard significance threshold of 0.05, the p-value of 0.1625 shows a lack of evidence to reject the null hypothesis, indicating that there may not be a meaningful relationship between the variables being studied.

In table 2, a significantly bigger test statistic (G = 572.68) with 1 degree of freedom resulted in an extremely small p-value (< 2.2e-16), showing a highly significant relationship. In this case, the evidence is overwhelming, resulting in the rejection of the null hypothesis.

## Contingency test for Independent Variables

```
# Create contingency table for the corresponding independent model
independent_table <-rbind(table(Both_eco_change$BD5up),table(Both_eco_change$BD11up))  # Assuming only BD11up is considered in the independent model
rownames(independent_table) <- c("BD5up","BD11up")
independent_table
```

```
##        Decrease Increase
## BD5up      1687      953
## BD11up     1638     1002
```

The above presented independent contingency table includes counts for the BD5up and BD11up categories.

## Odds Ratio, Sensitivity, Specificity & Youden's Index

```
contingency_observed <- Table_up_down

# Extract values from the contingency table
a <- contingency_observed[2, 2]  # True positives
b <- contingency_observed[1, 2]  # False negatives
c <- contingency_observed[2, 1]  # False positives
d <- contingency_observed[1, 1]  # True negatives

# Calculate odds ratio
odds_ratio <- (a * d) / (b * c)

# Calculate sensitivity
sensitivity <- a / (a + b)

# Calculate specificity
specificity <- d / (c + d)

# Calculate Youden's Index
youdens_index <- sensitivity + specificity - 1

# Display the calculated values
cat("Odds Ratio:", odds_ratio, "\n")
```

```
## Odds Ratio: 7.933389
```

```
cat("Sensitivity:", sensitivity, "\n")
```

```
## Sensitivity: 0.6789087
```

```
cat("Specificity:", specificity, "\n")
```

```
## Specificity: 0.7895673
```

```
cat("Youden's Index:", youdens_index, "\n")
```

```
## Youden's Index: 0.468476
```

The odds ratio indicates a strong relationship between BD11 and BD5 as they increase, and the sensitivity and specificity values indicate how well the model detects increases and decreases, respectively. Youden's Index evaluates both sensitivity and specificity to provide an overall assessment of the model's predictive ability. These values combined indicate a reasonable, but not perfect, link between BD11 and BD5 alterations.

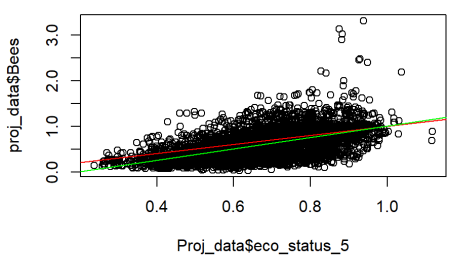
# Simple Linear Regression

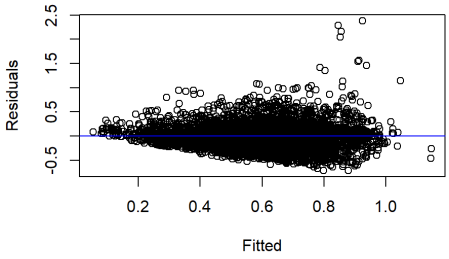
```
lin_mod <- lm(proj_data$Bees~Proj_data$eco_status_5)
# summary(lin_mod)

cat("Estimated Slope : ",coef(lin_mod)[2], "\n")
```

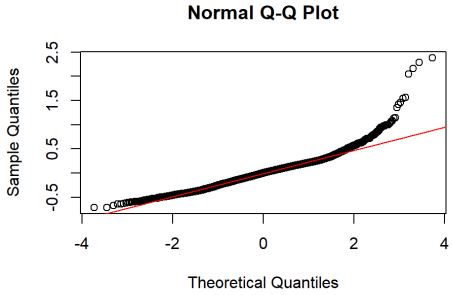
```
## Estimated Slope :  1.242784
```

A slope of about 1.242784 depicts that for every unit increase in the predictor variable, "eco_status_5," the response variable, "Bees," is expected to grow by approximately 1.242784 units on average.

```
plot(proj_data$Bees~Proj_data$eco_status_5)
abline(0,1,col="red")
abline(lin_mod,col="green")

plot(jitter(fitted(lin_mod)),residuals(lin_mod),xlab="Fitted",ylab="Residuals")
abline(h=0,col="blue")

qqnorm(residuals(lin_mod))
qqline(residuals(lin_mod),col="red")
```




From the above plot it can be easily observed that the data points appear to follow a linear pattern in the plot, with a rising trend when Proj_data$eco_status_5 rises. The positive slope of a green line is significantly different from 0, supports this association. From theoretical point of view both the variables should coincide perfectly along the red line with slope 1 and y-intercept 0. Due to some fluctuation or inconsistent(mistakes) in the data, there can be difference between the red line and the green line. Ultimately, the above figure gives a helpful visual representation between the two variables and how closely they are linked. In the second figure, the blue line, which roughly corresponds to the centre of the dots in the diagram. It depicts that the residuals seem to be symmetrically distributed around zero.In the third plot's appearance of a pretty straight line between the points shows that the residuals are likely to be regularly distributed. A reference line that shows how the points should look if they follow a normal distribution is the qqline plot. There is a red line that indicates that that the residuals are distributed regularly. Overall, the graph above demonstrates that the linear regression model according to testing meets the normality requirements.

# Multiple Linear Regression

## AIC for Initial MLR model

```
initial_MLR_model  <- lm(proj_data$Bees~.,
            data=Proj_data[c(my_variables)],y=TRUE)

# summary (initial_MLR_model )  # model summary
p_value <- max(summary(initial_MLR_model)$coefficients[, "Pr(>|t|)"])
p_value
```

```
## [1] 0.153108
```

```
AIC(initial_MLR_model )
```

```
## [1] 70.18494
```

The Initial MLR (Multiple Linear Regression) model has an AIC (Akaike Information Criterion) score of 70.18494 and has maximum p-value 0.153108, which indicates the model's quality of fit and complexity in explaining the variation in the data, taking into account the number of predictors and the overall performance.

## MLR Reduced AIC and Feature Selection by eliminating 'Carabids'

```
reduced_lmMod <- lm(proj_data$Bees~.,
                    data=Proj_data[c("Butterflies","Hoverflies","Ladybirds","Grasshoppers_._Crickets")],y=TRUE)
# summary(reduced_lmMod)
AIC(reduced_lmMod) # here lmMod is preferred by p and AIC criteria
```

```
## [1] 70.22846
```

```
summary (initial_MLR_model )
```

```
##
## Call:
## lm(formula = proj_data$Bees ~ ., data = Proj_data[c(my_variables)],
##     y = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71892 -0.16299  0.00213  0.14476  2.28485
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -0.45472    0.02402 -18.929  < 2e-16 ***
## Butterflies              0.65661    0.02692  24.387  < 2e-16 ***
## Carabids                 0.02972    0.02080   1.429    0.153
## Hoverflies               0.11705    0.02400   4.877 1.11e-06 ***
## Ladybirds                0.43973    0.01459  30.147  < 2e-16 ***
## Grasshoppers_._Crickets  0.18749    0.01915   9.792  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2434 on 5274 degrees of freedom
## Multiple R-squared:  0.3864, Adjusted R-squared:  0.3858
## F-statistic: 664.3 on 5 and 5274 DF,  p-value: < 2.2e-16
```

In the linear regression summary output, we can find the coefficients and p-values for each predictor in the model. The p-value for 'Carabids' is around 0.153. P-values reflect the relevance of each predictor in the model. Typically, if the p-value is less than a predetermined significance level (in this instance, 0.05), the predictor is considered statistically significant in explaining the variation in the response variable. In this situation, 'Carabids' has a p-value of 0.153, which exceeds 0.05. It implies that 'Carabids' might not be statistically significant in explaining the variation in 'proj_data$Bees'. So AIC for MLR Reduced model is 70.22846.

## AICs for MLR Interaction and MLR Reduced

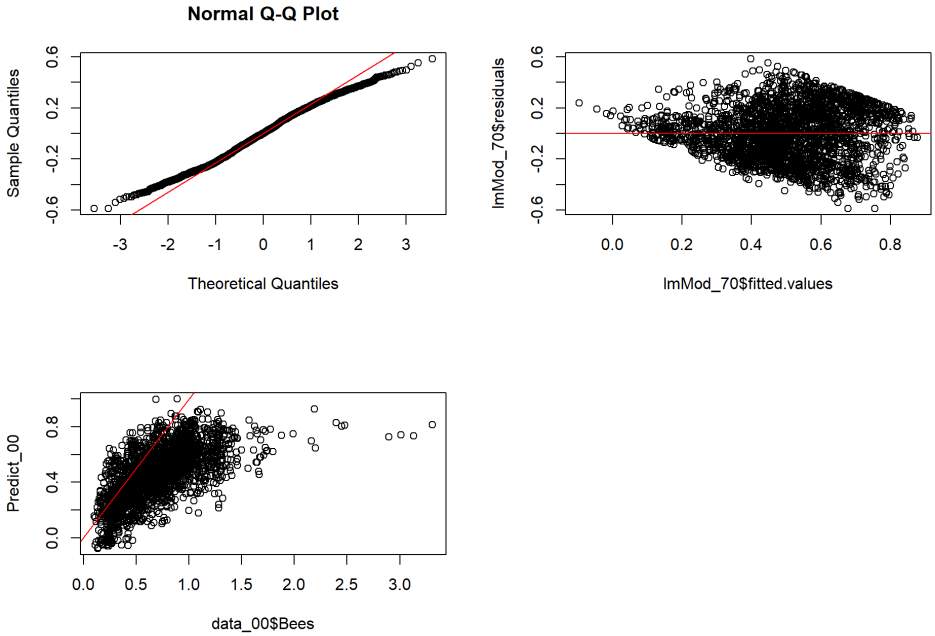
```
# now introduce an interaction
lmMod_interaction <- lm(proj_data$Bees~
                          Butterflies+Carabids+Hoverflies+Ladybirds+Grasshoppers_._Crickets
                        +Ladybirds*Carabids,
                        data=Proj_data,y=TRUE)
# summary(lmMod_interaction )
AIC(initial_MLR_model,reduced_lmMod,lmMod_interaction) # model with interaction preferred
```

```
##                   df      AIC
## initial_MLR_model  7 70.18494
## reduced_lmMod      6 70.22846
## lmMod_interaction  8 67.30563
```

```
# cor(lmMod_interaction$fitted.values,lmMod_interaction$y) # corelation slightly improved only
```

The Akaike Information Criterion (AIC) is a model selection measure that considers both quality of fit and model complexity. Lower AIC values imply a better balance between explaining the variation in the data and keeping the model simple. For `initial_MLR_model` With an AIC of around 70.18, this model strikes a reasonable compromise between quality of fit and model complexity. It's reasonably competitive, although there may be room for improvement. `reduced_lmMod` has a larger AIC value of 70.22846 compared to the baseline MLR model, indicating decreased performance or greater complexity. It might imply overfitting or insufficient explanatory power. Lastly for `lmMod_interaction` model has the lowest AIC of the three, suggesting a considerably better balance between goodness of fit and model complexity than the other models. It implies stronger performance in understanding the data, maybe with better predictions on fresh data.

```
# Extracting data of two different periods Y70 and Y00.
data_70 <- proj_data%>%filter(period=="Y70") # training set
data_00 <- proj_data%>%filter(period=="Y00") # test set
#nrow(data_00);nrow(data_00)

lmMod_70 <- lm(data_70$Bees~.,
               data=data_70[c(my_variables)],y=TRUE)
qqnorm(lmMod_70$residuals);qqline(lmMod_70$residuals,col="red")
plot(lmMod_70$residuals~lmMod_70$fitted.values) # look for unwanted pattern in residuals
abline(0,0,col="red")
Predict_00 <- predict(lmMod_70,data_00)
plot(Predict_00~data_00$Bees)
abline(0,1,col="red")
```




The above graph illustrates the relationship between the actual values of five taxonomic groups in the test data and proj_data$Bees as a anticipated values. The red line represents the ideal prediction line, which would have equal actual and predicted values. From the observations it appears like that the model fits the data well. In this particular instance, the residuals appear to be randomly distributed around the horizontal line at 0, indicating that the model fits the data well. The residuals from what is expected are distributed against a normal distribution in the qqplot demonstrated above. In a linear regression model, the residuals are assumed to have a normal distribution with a mean of 0 and a constant variance. Overall, the image indicates that the multiple linear regression model fits the data well and that the residuals normality assumptions are not considerably violated.

## Mean Square Errors for train and test data

```
cat("Mean Square Error on train data set:",mean((data_70$Bees-lmMod_70$fitted.values)^2),"\n")  # MSE on train data set
```

```
## Mean Square Error on train data set: 0.04094151
```

```
cat("Mean Square Error on test data set:",mean((data_00$Bees-Predict_00)^2),"\n")  # MSE on test data (higher)
```

```
## Mean Square Error on test data set: 0.1233863
```

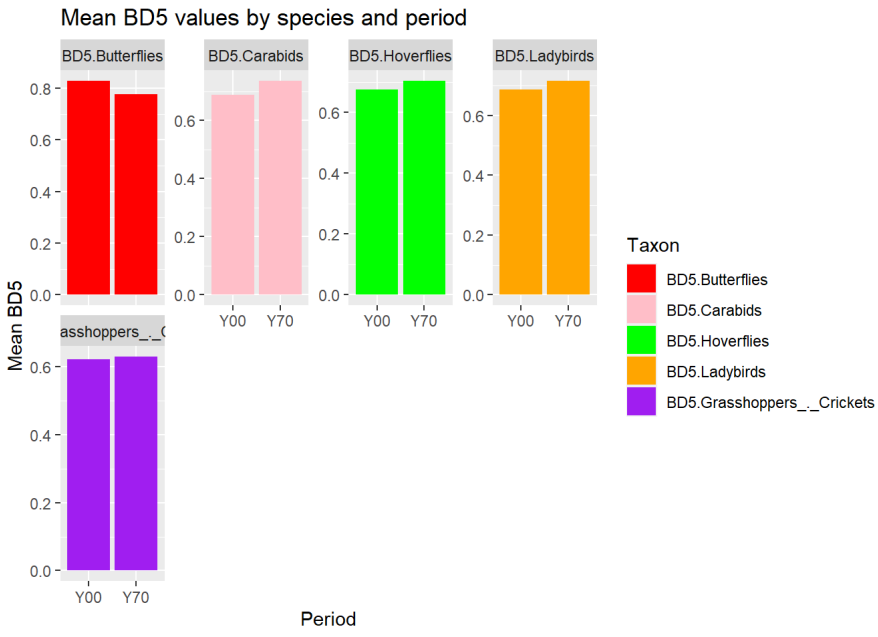
The MSE on the test set (Y00) is significantly greater than the MSE on the training set (Y70), indicating a performance gap between the two datasets. A significant difference between the two MSE values could indicate that the model is less effective when applied to the Y00 dataset, implying overfitting or a lack of generalizability to fresh data. While the model performs reasonably well on the training data (Y70) with a low MSE, it makes more errors when predicting outcomes in the test data (Y00), indicating potential problems with generalization or overfitting to the training data. Additional refinement or regularization approaches may be required to improve the model's capacity to generalize to new datasets.

# Open Analysis

```
# By using the aggregate function to calculate the mean of certain variables in the data frame proj_data based on period and
DominantLandClass

BD5_by_location_period <- aggregate(proj_data[my_variables], by = proj_data[c("dominantLandClass", "period")], FUN = mean)

# Filter for only the two periods of interest
BD5_Y70 <- BD5_by_location_period[BD5_by_location_period$period == "Y70", ]
BD5_Y00 <- BD5_by_location_period[BD5_by_location_period$period == "Y00", ]

# Merge the data with the corresponding land classes
land_classes <- data.frame(
  dominantLandClass = c("3e", "4e"),
  LandClass = c("Flat/gently undulating plains, E Anglia/S England","Flat coastal plains, E Anglia/S England")
)

BD5_Y70 <- merge(BD5_Y70, land_classes, by = "dominantLandClass")
BD5_Y00 <- merge(BD5_Y00, land_classes, by = "dominantLandClass")

# Create a new data frame with the mean BD7 values for each land class and period
BD5_means <- rbind(
  data.frame(LandClass = BD5_Y70$LandClass, BD5 = BD5_Y70[, my_variables], Period = "Y70"),
  data.frame(LandClass = BD5_Y00$LandClass, BD5 = BD5_Y00[, my_variables], Period = "Y00")
)

# Remove LandClass column from BD7_means
BD5_means <- subset(BD5_means, select = -LandClass)



# Melt data for plotting
BD5_means <- reshape2::melt(BD5_means, id.vars = c("Period"), variable.name = "Taxon", value.name = "Mean")
```

## Bar Plots

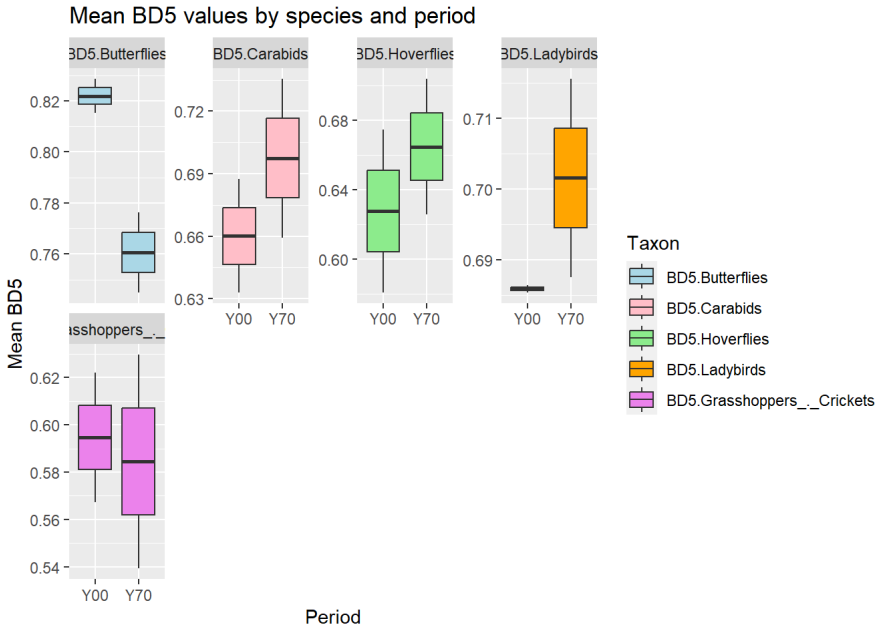
```
# Creating a bar plot of the mean BD5 values for each period

Bar_Plot<-ggplot(BD5_means, aes(x = Period, y = Mean, fill = Taxon)) +
  geom_col(position = "dodge") +
  ggtitle("Mean BD5 values by species and period") +
  xlab("Period") + ylab("Mean BD5") +
  facet_wrap(~Taxon, ncol = 4, scales="free_y") +
  scale_fill_manual(values = c("red", "pink", "green", "orange", "purple"))
Bar_Plot
```




The code first combines the time period and dominating land class of the specified five taxonomic groups BD5. A new data frame is created for charting with mean values for each land type and time period. With the help of ggplot it then generates the bar plot, where the bars are positioned side by side for each BD5 taxonomic group. These bars are represented by a distinct color and the bars shows the means of 5 taxonomic groups over two time periods (Y00 and Y70).

## Box Plots

```
# Creating a box plot of the mean BD5 values for each period
Box_Plot<- ggplot(BD5_means, aes(x = Period, y = Mean ,fill = Taxon)) +
  geom_boxplot() +
  ggtitle("Mean BD5 values by species and period") +
  xlab("Period") + ylab("Mean BD5") +
  facet_wrap(~Taxon, ncol = 4, scales = "free_y") +
  scale_fill_manual(values = c("lightblue", "pink", "lightgreen", "orange", "violet", "yellow","gray"))
Box_Plot
```




The graph above depicts the variation in BD5 levels over two unique time periods and species. We can also compare the median and range of BD5 values across species and time periods. Colours contribute in identifying and differentiation of diverse species. This map clearly depicts the distribution of BD5 mean values across species and time periods, allowing us to compare data and draw inferences.

# Conclusion

This report depicts and discusses how various plots were interpreted. The code computes and performs data exploration, summarize data (BD5), correlation between all pairs of variables of BD5. The program then performs hypotheses tests using T-Test and KS Test. Furthermore, the program performs Simple linear regression and multiple linear regression with qqplots respectively. Open Analysis is performed which gives Bar Plot and Box plot for all the BD5 variables over two time periods.

# References

1. https://moodle.essex.ac.uk/course/view.php?id=15074 (https://moodle.essex.ac.uk/course/view.php?id=15074)
2. https://moodle.essex.ac.uk/pluginfile.php/2016015/mod_resource/content/15/Guideline%20code%20for%20the%20MA334%20assignment%20A
(https://moodle.essex.ac.uk/pluginfile.php/2016015/mod_resource/content/15/Guideline%20code%20for%20the%20MA334%20assignment%20A