# Ocean Data Analysis with R Programming for Early Career Ocean Professionals (ECOPs) (Asia)

## Module 2 - Lesson 2: Exploring the Relationship between Environmental Factors and Marine Species Distribution

Mohamad Lukman Aidid bin Mohd Yusoff

2023-10-21

*Project I: Plotting the Distribution of Marine Species along Shore Distance and Depth*

**Task:**

1. Use ggplot in R to create multiple plots to explore the relationship between environmental factors (sst, sss, minimum depth in meters, and shore distance) and marine species distribution (number of individuals, family, and Red List category).
2. Fit a multiple regression model to predict the number of individuals based on the environmental factors.
3. Evaluate the model fit using appropriate metrics (e.g., R^2, AIC, BIC, residual plots) and interpret the results.
4. Provide a brief summary of the results and recommendations for future research based on the findings.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(patchwork)
```

```
obis = read.csv(file = "obis_red_list_filtered_1000.csv")
head(obis)
```

```
##          scientificName date_year          family minimumDepthInMeters
## 1 Balaenoptera physalus      2003 Balaenopteridae                     0
## 2 Balaenoptera physalus      2003 Balaenopteridae                     0
## 3 Balaenoptera physalus      2003 Balaenopteridae                     0
## 4 Balaenoptera physalus      2003 Balaenopteridae                     0
```

```
## 5 Balaenoptera physalus      2003 Balaenopteridae                    0
## 6 Balaenoptera physalus      2002 Balaenopteridae                    0
##   shoredistance   sst   sss individualCount    country status
## 1        182964 -1.47 34.03               2 Antarctica     VU
## 2        135623 -1.58 34.01               2 Antarctica     VU
## 3        138638 -1.58 34.01               9 Antarctica     VU
## 4         77966 -1.57 34.06               4 Antarctica     VU
## 5        141441 -1.59 34.02               3 Antarctica     VU
## 6        -14124 -1.43 33.71               3 Antarctica     VU
```

```
str(obis)
```

```
## 'data.frame':    1000 obs. of  10 variables:
##  $ scientificName      : chr  "Balaenoptera physalus" "Balaenoptera physalus" "Balaenoptera physalus
##  $ date_year           : int  2003 2003 2003 2003 2003 2002 2003 2003 2003 2020 ...
##  $ family              : chr  "Balaenopteridae" "Balaenopteridae" "Balaenopteridae" "Balaenopteridae
##  $ minimumDepthInMeters: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ shoredistance       : int  182964 135623 138638 77966 141441 -14124 727065 184171 144748 478287 .
##  $ sst                 : num  -1.47 -1.58 -1.58 -1.57 -1.59 -1.43 -0.51 -1.48 -1.55 0.35 ...
##  $ sss                 : num  34 34 34 34.1 34 ...
##  $ individualCount     : num  2 2 9 4 3 3 3 6 6 8 ...
##  $ country             : chr  "Antarctica" "Antarctica" "Antarctica" "Antarctica" ...
##  $ status              : chr  "VU" "VU" "VU" "VU" ...
```

```
summary(obis)
```

```
##  scientificName       date_year       family          minimumDepthInMeters
##  Length:1000        Min.   :2000   Length:1000        Min.   :   0.00
##  Class :character   1st Qu.:2003   Class :character   1st Qu.:   4.00
##  Mode  :character   Median :2006   Mode  :character   Median :  12.50
##                     Mean   :2008                      Mean   :  91.41
##                     3rd Qu.:2015                      3rd Qu.:  56.00
##                     Max.   :2020                      Max.   :1346.00
##  shoredistance            sst             sss        individualCount
##  Min.   : -14124.0   Min.   :-1.63   Min.   :21.82   Min.   :   1.177
##  1st Qu.:    670.8   1st Qu.:16.54   1st Qu.:33.78   1st Qu.:   2.000
##  Median :   5664.0   Median :21.84   Median :34.86   Median :   3.000
##  Mean   :  95059.8   Mean   :20.50   Mean   :34.44   Mean   :  21.114
##  3rd Qu.:  27220.0   3rd Qu.:27.82   3rd Qu.:35.59   3rd Qu.:   6.000
##  Max.   :1775379.0   Max.   :29.81   Max.   :37.38   Max.   :2796.000
##    country             status
##  Length:1000        Length:1000
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

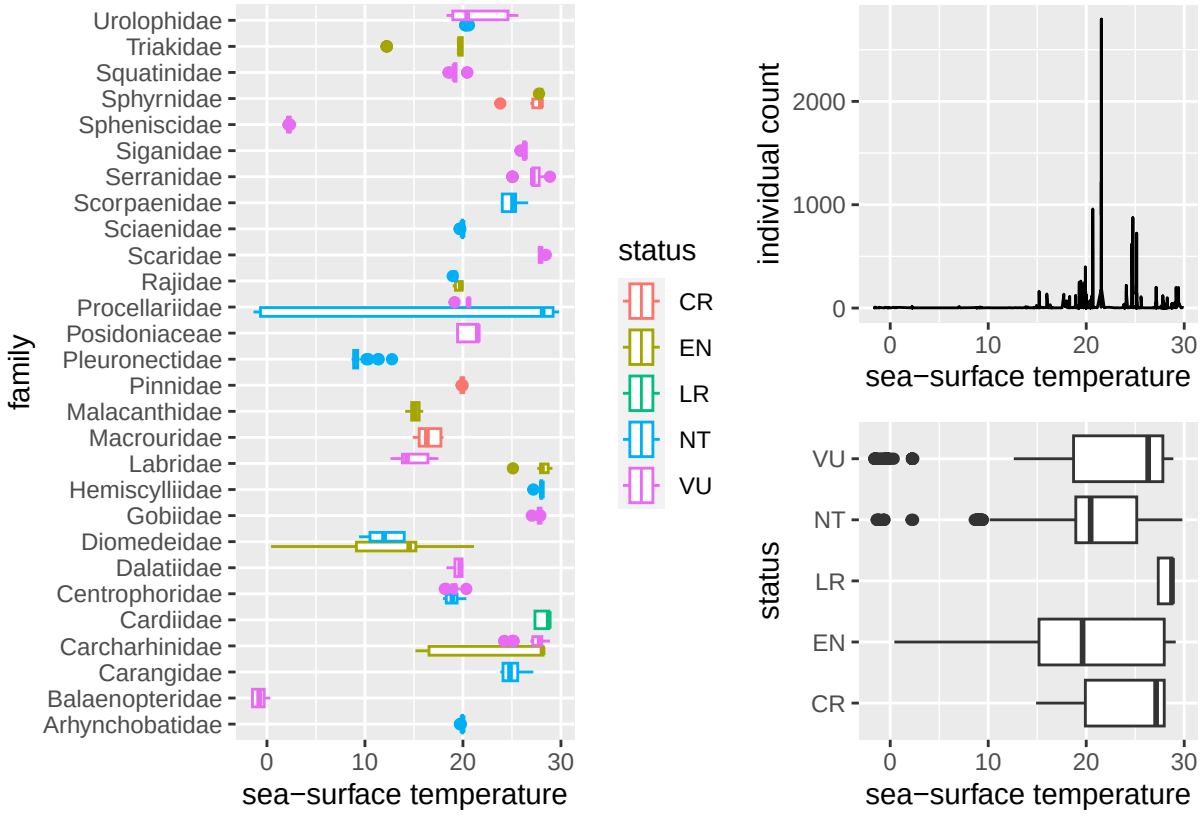
Plots of Sea-Surface Temperature vs Marine Species Distribution

```
plot_sst_count <- ggplot(data = obis, aes(x = sst, y = individualCount)) +
  geom_line() +
  labs(x = "sea-surface temperature", y = "individual count")

plot_sst_family <- ggplot(data = obis, aes(x = sst, y = family, color=status)) +
  geom_boxplot() +
  labs(x = "sea-surface temperature", y = "family")

plot_sst_status <- ggplot(data = obis, aes(x = sst, y = status)) +
  geom_boxplot() +
  labs(x = "sea-surface temperature", y = "status")

plot_sst_family + plot_sst_count / plot_sst_status
```



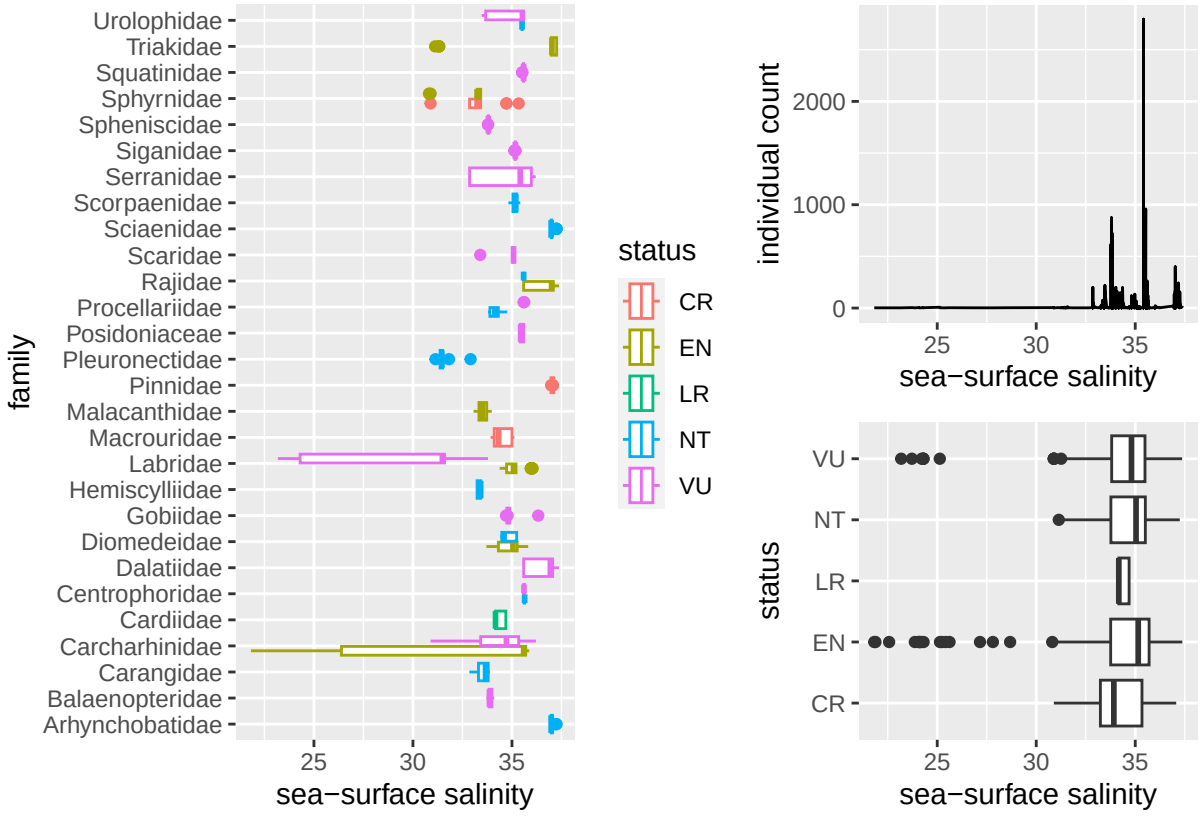
Plots of Sea-Surface Salinity vs Marine Species Distribution

```
plot_sss_count <- ggplot(data = obis, aes(x = sss, y = individualCount)) +
  geom_line() +
  labs(x = "sea-surface salinity", y = "individual count")

plot_sss_family <- ggplot(data = obis, aes(x = sss, y = family, color=status)) +
  geom_boxplot() +
  labs(x = "sea-surface salinity", y = "family")

plot_sss_status <- ggplot(data = obis, aes(x = sss, y = status)) +
  geom_boxplot() +
  labs(x = "sea-surface salinity", y = "status")

plot_sss_family  + plot_sss_count / plot_sss_status
```



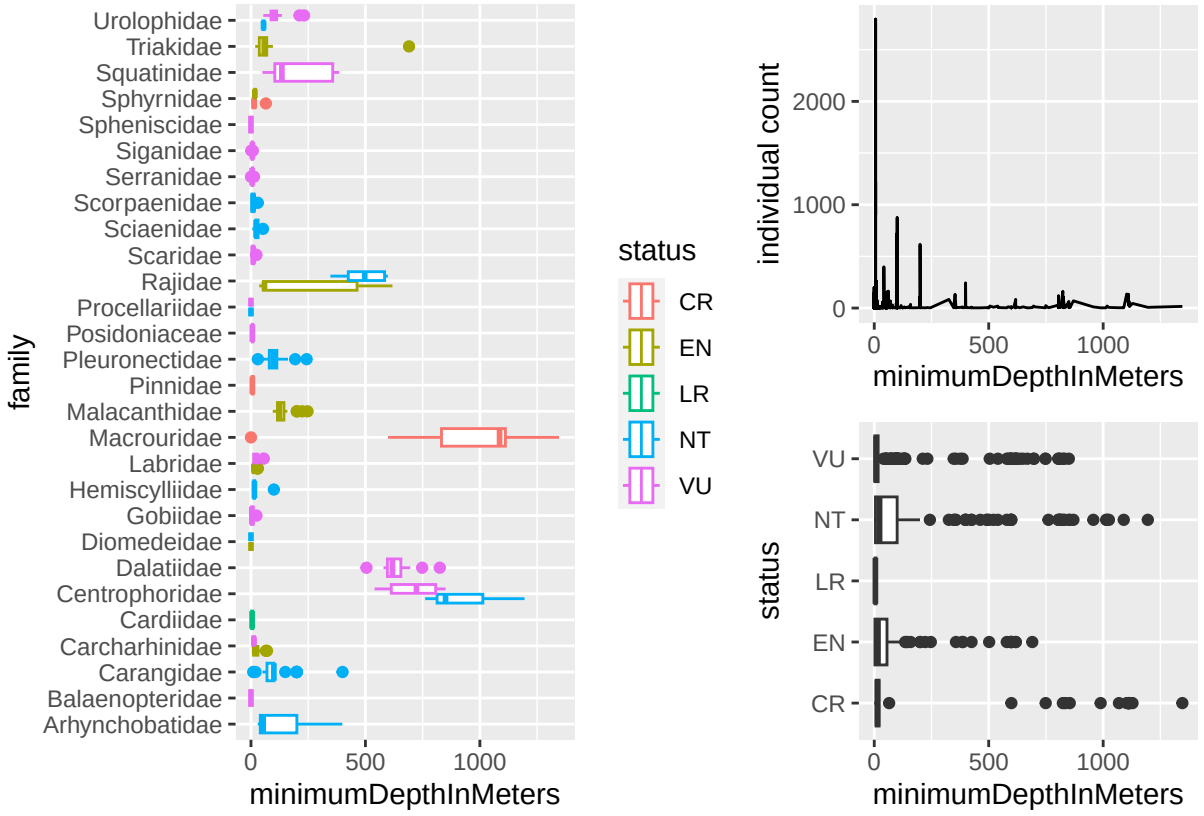
Plots of Minimum Depth vs Marine Species Distribution

```
plot_depth_count <- ggplot(data = obis, aes(x = minimumDepthInMeters, y = individualCount)) +
  geom_line() +
  labs(x = "minimumDepthInMeters", y = "individual count")

plot_depth_family <- ggplot(data = obis, aes(x = minimumDepthInMeters, y = family, color=status)) +
  geom_boxplot() +
  labs(x = "minimumDepthInMeters", y = "family")

plot_depth_status <- ggplot(data = obis, aes(x = minimumDepthInMeters, y = status)) +
  geom_boxplot() +
  labs(x = "minimumDepthInMeters", y = "status")

plot_depth_family + plot_depth_count / plot_depth_status
```



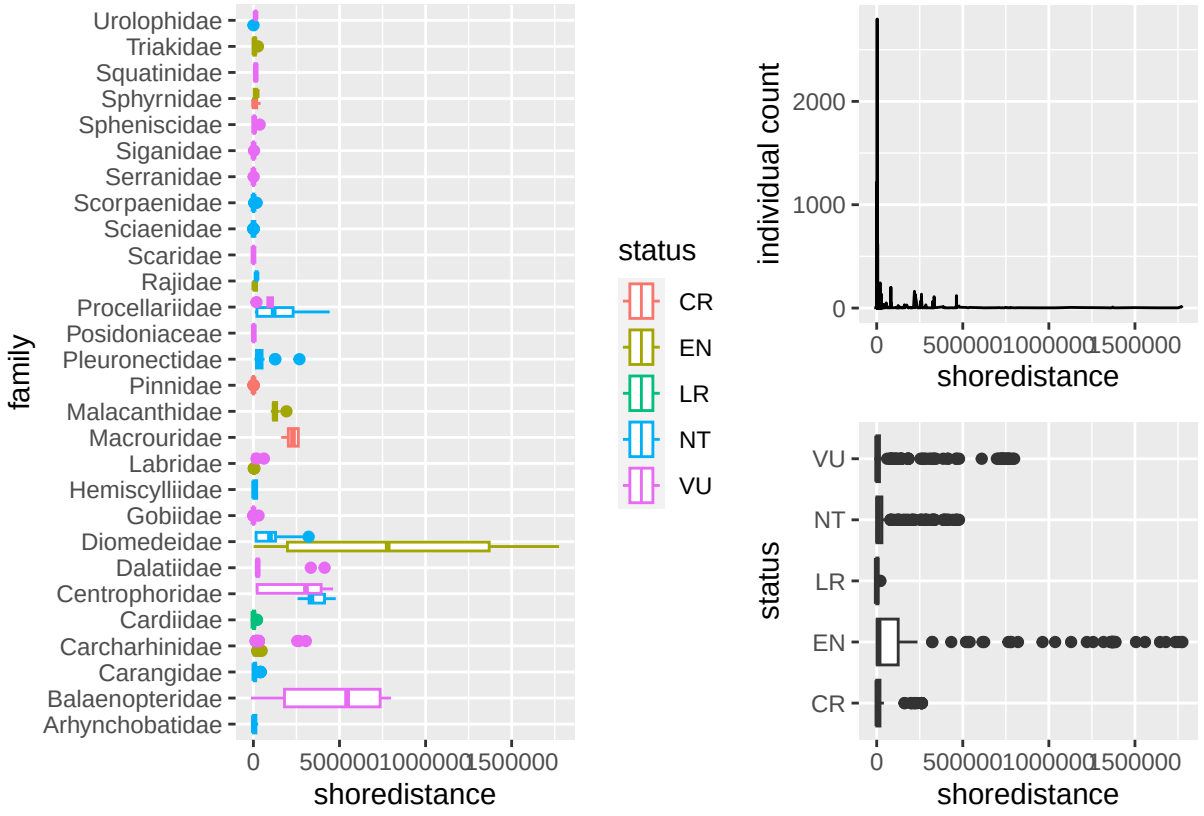
Plots of Shore Distance vs Marine Species Distribution

```
plot_shoredistance_count <- ggplot(data = obis, aes(x = shoredistance, y = individualCount)) +
  geom_line() +
  labs(x = "shoredistance", y = "individual count")

plot_shoredistance_family <- ggplot(data = obis, aes(x = shoredistance, y = family, color=status)) +
  geom_boxplot() +
  labs(x = "shoredistance", y = "family")

plot_shoredistance_status <- ggplot(data = obis, aes(x = shoredistance, y = status)) +
  geom_boxplot() +
  labs(x = "shoredistance", y = "status")

plot_shoredistance_family + plot_shoredistance_count / plot_shoredistance_status
```



Modeling using glm() function

```
model = glm(individualCount ~ sst + sss + minimumDepthInMeters + shoredistance, data = obis)

summary(model)
```

```
##
## Call:
## glm(formula = individualCount ~ sst + sss + minimumDepthInMeters +
##     shoredistance, data = obis)
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -7.729e+01  6.490e+01  -1.191    0.234
## sst                   8.118e-02  5.260e-01   0.154    0.877
## sss                   2.843e+00  1.938e+00   1.467    0.143
## minimumDepthInMeters  6.169e-03  1.837e-02   0.336    0.737
## shoredistance        -1.818e-05  1.661e-05  -1.095    0.274
##
## (Dispersion parameter for gaussian family taken to be 14788.95)
##
##     Null deviance: 14778412  on 999  degrees of freedom
## Residual deviance: 14715008  on 995  degrees of freedom
## AIC: 12447
##
## Number of Fisher Scoring iterations: 2
```

Modeling using lm() function

```
model2 = lm(individualCount ~ sst + sss + minimumDepthInMeters + shoredistance, data = obis)

summary(model2)
```

```
##
## Call:
## lm(formula = individualCount ~ sst + sss + minimumDepthInMeters +
##     shoredistance, data = obis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -31.18  -22.58  -17.23   -9.55 2770.85
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -7.729e+01  6.490e+01  -1.191    0.234
## sst                   8.118e-02  5.260e-01   0.154    0.877
## sss                   2.843e+00  1.938e+00   1.467    0.143
## minimumDepthInMeters  6.169e-03  1.837e-02   0.336    0.737
## shoredistance        -1.818e-05  1.661e-05  -1.095    0.274
##
## Residual standard error: 121.6 on 995 degrees of freedom
## Multiple R-squared:  0.00429,    Adjusted R-squared:  0.0002875
## F-statistic: 1.072 on 4 and 995 DF,  p-value: 0.3691
```

**Conclusion:**

- For this particular dataset, crude Gaussian-based multiple regression on the four environmental factors (namely 'sea-surface temperature', 'sea-surface salinity', 'minimum depth', and 'shore distance') *do not* give a very good modeling of 'individual count'.
- More sophisticated tests and modeling are needed to properly model the dataset.