## Load and Clean Datasets

### 1. Monitoring Data

Loads RB5 monitoring data, filters for the target period, and drops unused variables.

RB5\_monitoring\_data <- read.csv("../data/csv/RB5\_monitoring\_data\_monthly\_average\_1980-2023.csv")  
RB5\_monitoring\_data <- RB5\_monitoring\_data %>%  
 filter(between(year, 2004, 2016)) %>%  
 select(-RB5.Cond, -RB5.NO2, -RB5.NH4, -RB5.PP, -RB5.PSiO2, -RB5.SURP, -RB5.TSi, -RB5.TSP, -RB5.TON) %>%  
 mutate(YearMonth = as.Date(YearMonth))

### 2. Biovolume Data

Processes monthly biovolume values and standardizes date format.

biovol\_fulldata <- read.csv("../data/csv/biovolfulldata\_2004-2016.csv") %>%  
 mutate(YearMonth = as.Date(YearMonth))

### 3. Predator (Daphnia) Data

Selects and processes zooplankton data, aggregates monthly means, and filters for the target timeframe.

predator\_data <- read.csv("../data/csv/crustaceanzooplankton\_1980-2023.csv") %>%  
 select(Date, Daphnia.hyalina.ind.L) %>%  
 mutate(Date = dmy(Date), year = year(Date), month = month(Date)) %>%  
 filter(!is.na(Date)) %>%  
 mutate(Daphnia = as.numeric(Daphnia.hyalina.ind.L)) %>%  
 select(-Daphnia.hyalina.ind.L) %>%  
 group\_by(year, month) %>%  
 summarise(across(where(is.numeric), mean, na.rm = TRUE), .groups = 'drop') %>%  
 mutate(YearMonth = as.Date(paste(year, month, "01", sep = "-"))) %>%  
 filter(between(YearMonth, as.Date("2004-01-01"), as.Date("2016-12-01")))

### 4. Water Level Data

Combines two water level datasets and computes monthly means.

waterlevelp2 <- read.csv("../data/csv/waterlevels\_1993-2007.csv") %>%  
 rename(waterlevel.masl = Harbour.masl) %>%  
 mutate(Date = dmy(Date)) %>%  
 select(Date, waterlevel.masl)  
  
waterlevelp3 <- read.csv("../data/csv/waterlevels\_2008-2013.csv") %>%  
 rename(Date = SAMPLE\_DATE, waterlevel.masl = Level.maod) %>%  
 mutate(Date = dmy(Date)) %>%  
 select(Date, waterlevel.masl)  
  
fullwaterlevel <- bind\_rows(waterlevelp2, waterlevelp3) %>%  
 mutate(year = year(Date), month = month(Date)) %>%  
 group\_by(year, month) %>%  
 summarise(across(where(is.numeric), mean, na.rm = TRUE), .groups = 'drop') %>%  
 mutate(YearMonth = as.Date(paste(year, month, "01", sep = "-"))) %>%  
 filter(between(year, 2004, 2016))

### 5. Temperature and SD Data

Processes temperature and standard deviation data, summarising to monthly means.

tempsddata <- read.csv("../data/csv/Temp\_SD\_2004-2016.csv") %>%  
 mutate(Date = dmy(Date), RB5.SD = as.numeric(RB5.SD), RB5.Temp = as.numeric(RB5.Temp),  
 year = year(Date), month = month(Date)) %>%  
 group\_by(year, month) %>%  
 summarise(across(where(is.numeric), mean, na.rm = TRUE), .groups = 'drop') %>%  
 mutate(YearMonth = as.Date(paste(year, month, "01", sep = "-")))

### 6. pH Data (2010–2016)

Processes pH values, available only for 2010–2016, into monthly means.

pHdata <- read.csv("../data/csv/pH\_2010-2016.csv") %>%  
 select(Date, Value) %>%  
 mutate(Date = dmy(Date), RB5.pH = Value,  
 year = year(Date), month = month(Date)) %>%  
 group\_by(year, month) %>%  
 summarise(across(where(is.numeric), mean, na.rm = TRUE), .groups = 'drop') %>%  
 mutate(YearMonth = as.Date(paste(year, month, "01", sep = "-")))

### 7. DO Data (2010–2016)

Filters and processes dissolved oxygen (% saturation) into monthly averages.

DOdata <- read.csv("../data/csv/DO\_2010-2016.csv") %>%  
 filter(Units == "%") %>%  
 select(Date, Value) %>%  
 mutate(Date = dmy(Date), RB5.DO = Value,  
 year = year(Date), month = month(Date)) %>%  
 group\_by(year, month) %>%  
 summarise(across(where(is.numeric), mean, na.rm = TRUE), .groups = 'drop') %>%  
 mutate(YearMonth = as.Date(paste(year, month, "01", sep = "-")))

## Merge All Data

This step incrementally merges all cleaned datasets into one by matching on the YearMonth column.

all\_data1 <- full\_join(RB5\_monitoring\_data, select(predator\_data, YearMonth, Daphnia), by = "YearMonth")  
all\_data2 <- full\_join(biovol\_fulldata, select(fullwaterlevel, YearMonth, waterlevel.masl), by = "YearMonth")  
all\_data <- full\_join(all\_data2, all\_data1, by = "YearMonth")  
  
all\_data <- all\_data %>%  
 left\_join(tempsddata, by = "YearMonth") %>%  
 mutate(RB5.SD = coalesce(RB5.SD.x, RB5.SD.y),  
 RB5.Temp = coalesce(RB5.Temp.x, RB5.Temp.y)) %>%  
 select(-RB5.SD.x, -RB5.SD.y, -RB5.Temp.x, -RB5.Temp.y) %>%  
 left\_join(pHdata, by = "YearMonth") %>%  
 mutate(RB5.pH = coalesce(RB5.pH.x, RB5.pH.y)) %>%  
 select(-RB5.pH.x, -RB5.pH.y) %>%  
 left\_join(DOdata, by = "YearMonth") %>%  
 mutate(RB5.DO = coalesce(RB5.DO.x, RB5.DO.y)) %>%  
 select(-RB5.DO.x, -RB5.DO.y)

## Final Cleaning

Refines merged dataset by selecting relevant columns, deriving seasonal categories, and calculating total biovolume.

all\_data <- all\_data %>%  
 distinct(YearMonth, .keep\_all = TRUE) %>%  
 select(YearMonth, year.y, month.y, Cryto.Biovolume, Cyano.Biovolume,  
 Diatoms.Biovolume, Greens.Biovolume, waterlevel.masl, Daphnia,  
 starts\_with("RB5")) %>%  
 rename(Year = year.y, Month = month.y) %>%  
 mutate(Season = case\_when(  
 Month %in% c(12, 1, 2) ~ "Winter",  
 Month %in% c(3, 4, 5) ~ "Spring",  
 Month %in% c(6, 7, 8) ~ "Summer",  
 Month %in% c(9, 10, 11) ~ "Autumn"  
 )) %>%  
 mutate(Total.Biovolume = Cryto.Biovolume + Cyano.Biovolume +  
 Diatoms.Biovolume + Greens.Biovolume) %>%  
 relocate(Season, .after = Month) %>%  
 relocate(Total.Biovolume, .after = Greens.Biovolume)

## Save Final Dataset

Exports the cleaned and merged dataset as CSV files for further analysis.

write.csv(all\_data, "../data/csv/alldata-2004-2016v2.csv", row.names = FALSE)