Data cleaning

2026-02-23

Data cleaning for initial comparison of reproductive output between sites.

This document outlines how the raw data was processed. The output is two plots. The first shows a comparison of sandprawn (Kraussillichirus kraussi) egg biomass across the five sites that were sampled. The second is a comparison of sandprawn embryo numbers across the sites.

Reading in the data

```
library(readxl)
rawdata <- read_excel("Ella sandprawn data.xlsx")
head(rawdata)
tail(rawdata)
```

Selecting the relevant coloumns: Site, Number of embryos and Egg biomass

```
spdata <- rawdata[, c(3, 10, 11)]
head(spdata)
```

```
## # A tibble: 6 x 3
##    Site `No. of embryos` `Egg biomass (g)`
##   <dbl>            <dbl>              <dbl>
## 1     1                0                  0
## 2     1                0                  0
## 3     1                0                  0
## 4     1                0                  0
## 5     1                0                  0
## 6     1                0                  0
```

```
str(spdata)
```

```
## tibble [789 x 3] (S3: tbl_df/tbl/data.frame)
##  $ Site           : num [1:789] 1 1 1 1 1 1 1 1 1 1 ...
##  $ No. of embryos : num [1:789] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Egg biomass (g): num [1:789] 0 0 0 0 0 0 0 0 0 0 ...
```

Converting Site from numeric to a factor

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.2.0     v readr     2.2.0
## v forcats   1.0.1     v stringr   1.6.0
## v ggplot2   4.0.2     v tibble    3.3.1
## v lubridate 1.9.5     v tidyr     1.3.2
## v purrr     1.2.1
## -- 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
```

```r
spdata <- spdata %>% mutate(Site= as.factor(Site))
str(spdata)
```

```
## tibble [789 x 3] (S3: tbl_df/tbl/data.frame)
##  $ Site           : Factor w/ 5 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ No. of embryos : num [1:789] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Egg biomass (g): num [1:789] 0 0 0 0 0 0 0 0 0 0 ...
```

Removing observations where animals did not have eggs

```r
clean_spdata <- spdata %>% filter(`Egg biomass (g)` > 0)
head(clean_spdata)
```

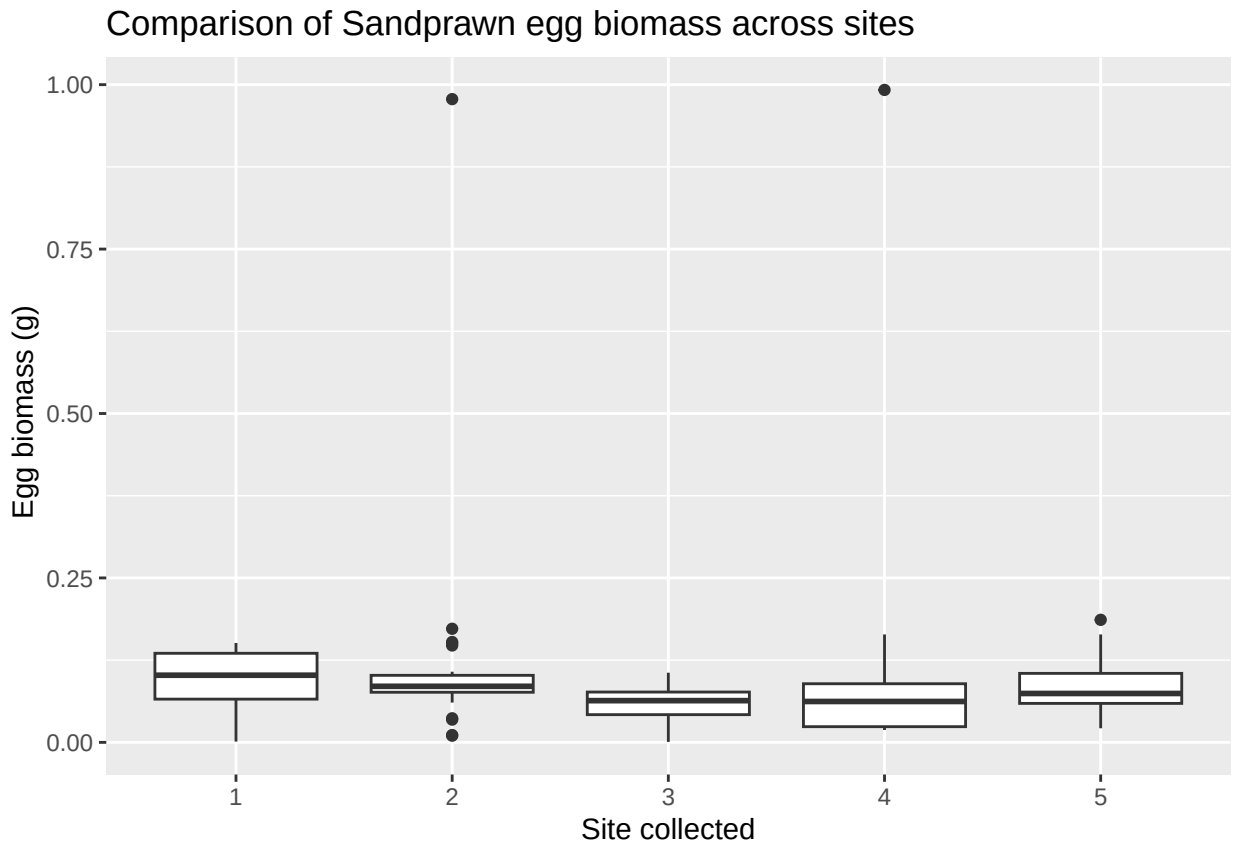
```
## # A tibble: 6 x 3
##   Site  `No. of embryos` `Egg biomass (g)`
##   <fct>            <dbl>             <dbl>
## 1 1                  117             0.124
## 2 1                   77             0.151
## 3 1                    1             0.0012
## 4 1                   90             0.141
## 5 1                   98             0.135
## 6 1                   74             0.132
```

```r
summary(clean_spdata)
```

```
##  Site   No. of embryos   Egg biomass (g)
##  1:17   Min.   :  1.00   Min.   :0.00070
##  2:28   1st Qu.: 43.25   1st Qu.:0.04398
##  3:29   Median : 64.00   Median :0.07680
##  4:22   Mean   : 63.57   Mean   :0.09014
##  5:30   3rd Qu.: 85.75   3rd Qu.:0.09978
##         Max.   :161.00   Max.   :0.99200
```

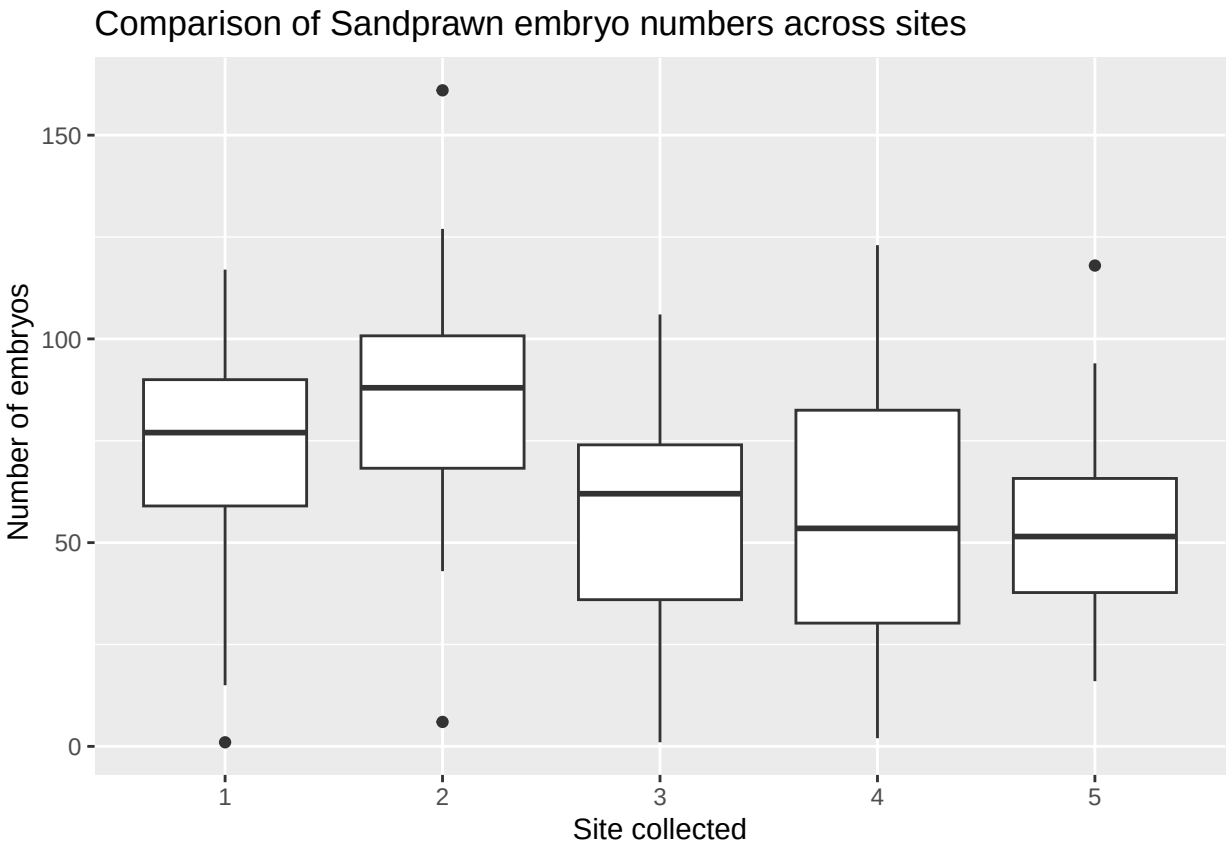
Creating boxplot of sandprawn egg biomass across sites

```
library(ggplot2)
ggplot(data = clean_spdata, aes(x = Site, y = `Egg biomass (g)`)) + geom_boxplot() + labs(title = "Compa
```



Creating boxplot of sandprawn embryo numbers across sites

```
ggplot(data = clean_spdata, aes(x = Site, y = `No. of embryos`)) + geom_boxplot() + labs(title = "Compa
```


Comparison of Sandprawn embryo numbers across sites
Number of embryos
150
100
50
0
1
2
3
4
5
Site collected