Green crab data processing

Data read in

These are the data sets, filtering steps, and libraries used to construct plots and statistical analyses. The main data cleaning step involved changing variable types (changing categorical factors to numeric, for example).

```
# read in relevant documents for this data
# preliminary data used for calcium etc.

# data used to produce calcium plots
# calcium data came from a limited set of crabs
gc_dat <- read.csv("../data/prelim_gc_data.csv")

# updated with full data set 2023-01-06

gc_dat_fin <- read.csv("../data/Condo_Data_Raw_CML_WEL_1_6_23.csv")

# define color levels as factor for future analyses
gc_dat_fin$hemo_col <- factor(gc_dat_fin$hemo_col,
                              levels = c('1', '2', '3', '4', '5'))

gc_dat_fin$hemo_col <- as.factor(gc_dat_fin$hemo_col)

# define hemolymph refractive index (RI) as numerics
gc_dat_fin$hemo_ri <- as.numeric(gc_dat_fin$hemo_ri)

# data reclassification to avoid errors later

# libraries for data visualization
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```r
library(janitor)
```

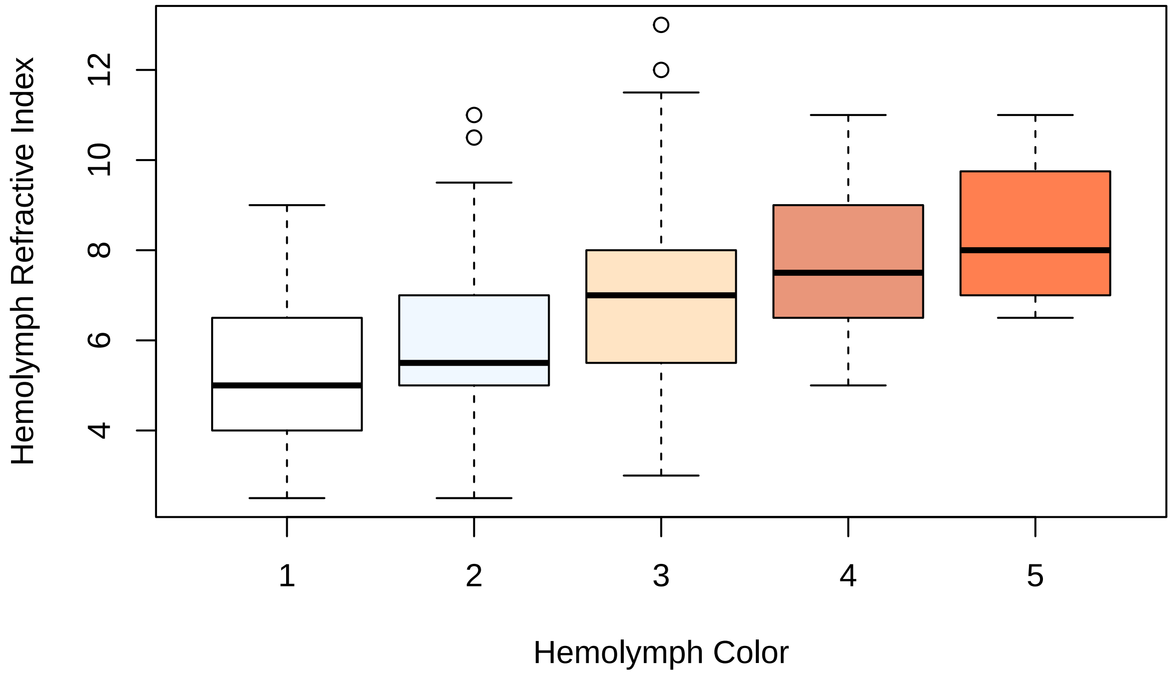
```
##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

RI vs. Color Plots

```r
boxplot(hemo_ri ~ hemo_col, data = gc_dat_fin,
      xlab = "Hemolymph Color",
      ylab = "Hemolymph Refractive Index",
      col= c('white', 'aliceblue', 'bisque', 'darksalmon','coral'),
      subset = condo_id == "WL" &
        outcome == c('Molted', 'Removed'))
```

```
## Warning in outcome == c("Molted", "Removed"): longer object length is not a
## multiple of shorter object length
```



```
fcol <- factor(gc_dat_fin$hemo_col)
# color is on a scale of 1-5; appears as "numeric"; categorical factor
# stores the variable hemolymph color as a factor (fcol)

ri_col_aov <- aov (hemo_ri ~ fcol, data=gc_dat_fin,
                   subset = condo_id == "WL" &
                     outcome == c('Molted', 'Removed'))
```

```
## Warning in outcome == c("Molted", "Removed"): longer object length is not a
## multiple of shorter object length
```

```
summary(ri_col_aov)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## fcol          4  219.5   54.88   16.75 1.77e-12 ***
## Residuals   319 1044.9    3.28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 266 observations deleted due to missingness
```

```
# type 1 aov used for Tukey test
# tests between categories: is 1 dif. than 2, etc.

TukeyHSD(ri_col_aov, conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = hemo_ri ~ fcol, data = gc_dat_fin, subset = condo_id == "WL" & outcome == c("Molte
##
## $fcol
##          diff        lwr      upr     p adj
## 2-1 0.6217576 -0.1444341 1.387949 0.1726677
## 3-1 1.6082925  0.9085657 2.308019 0.0000000
## 4-1 2.4842226  1.2713944 3.697051 0.0000004
## 5-1 3.1032702  1.1490905 5.057450 0.0001726
## 3-2 0.9865349  0.2918199 1.681250 0.0011217
## 4-2 1.8624650  0.6525213 3.072409 0.0003028
## 5-2 2.4815126  0.5291218 4.433903 0.0050213
## 4-3 0.8759301 -0.2930573 2.044917 0.2421847
## 5-3 1.4949777 -0.4322996 3.422255 0.2107252
## 5-4 0.6190476 -1.5479162 2.786011 0.9352081
```

New data frame to check previous results

To make sure that everything was showing up how I intended in the other figures and to double check the subsetting, I made a new data frame titled *filt* with only crabs from wells that molted and were removed. This data frame only contains the crab id, hemolymph ri, and hemolymph colors. I reconstructed the same boxplot and anova/tukey hsd testing.

```
gc_dat_clean <- clean_names(gc_dat_fin)
# use janitor to clean

filt <- gc_dat_clean[gc_dat_clean$condo_id == "WL" &
                     gc_dat_clean$outcome == c('Molted','Removed'),
                     c("crab_id", "hemo_ri", "hemo_col")]
```

```
## Warning in gc_dat_clean$outcome == c("Molted", "Removed"): longer object length
## is not a multiple of shorter object length
```

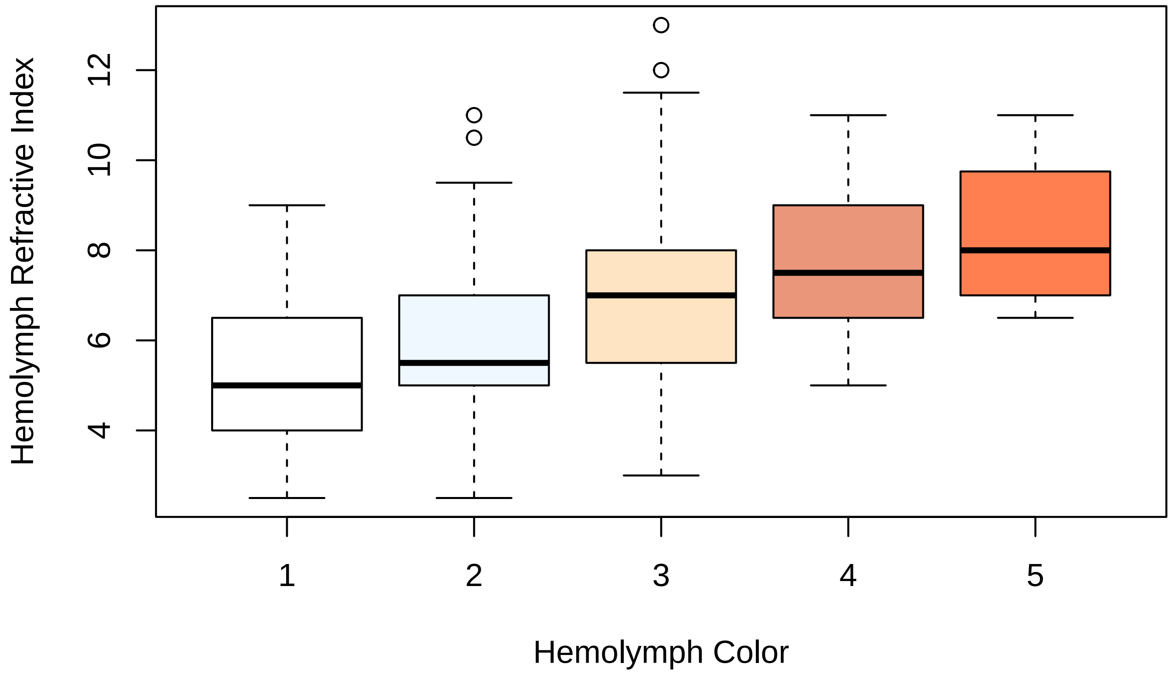
```
# write data table with filtered data

write.csv(filt, "filt_gc_dat.csv",
          sep="\t",
          quote=F,
          row.names = F,
          col.names = T)
```

```
## Warning in write.csv(filt, "filt_gc_dat.csv", sep = "\t", quote = F, row.names
## = F, : attempt to set 'col.names' ignored
```

```
## Warning in write.csv(filt, "filt_gc_dat.csv", sep = "\t", quote = F, row.names
## = F, : attempt to set 'sep' ignored
```

```
# make boxplot with filtered data to double check other graph
boxplot(hemo_ri ~ hemo_col, data = filt,
      xlab = "Hemolymph Color",
      ylab = "Hemolymph Refractive Index",
      col= c('white', 'aliceblue', 'bisque', 'darksalmon','coral'))
```



```
# n = 83, 85, 125, 21, 7
# gc n = 92

fcol_filt <- factor(filt$hemo_col)
# color is on a scale of 1-5; appears as "numeric"; categorical factor

ri_col_aov_filt <- aov (hemo_ri ~ fcol_filt, data=filt)
summary(ri_col_aov_filt)
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## fcol_filt       4  219.5   54.88   16.75 1.77e-12 ***
## Residuals     319 1044.9    3.28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 266 observations deleted due to missingness
```

```
# type 1 aov used for Tukey test
# tests between categories: is 1 dif. than 2, etc.
```

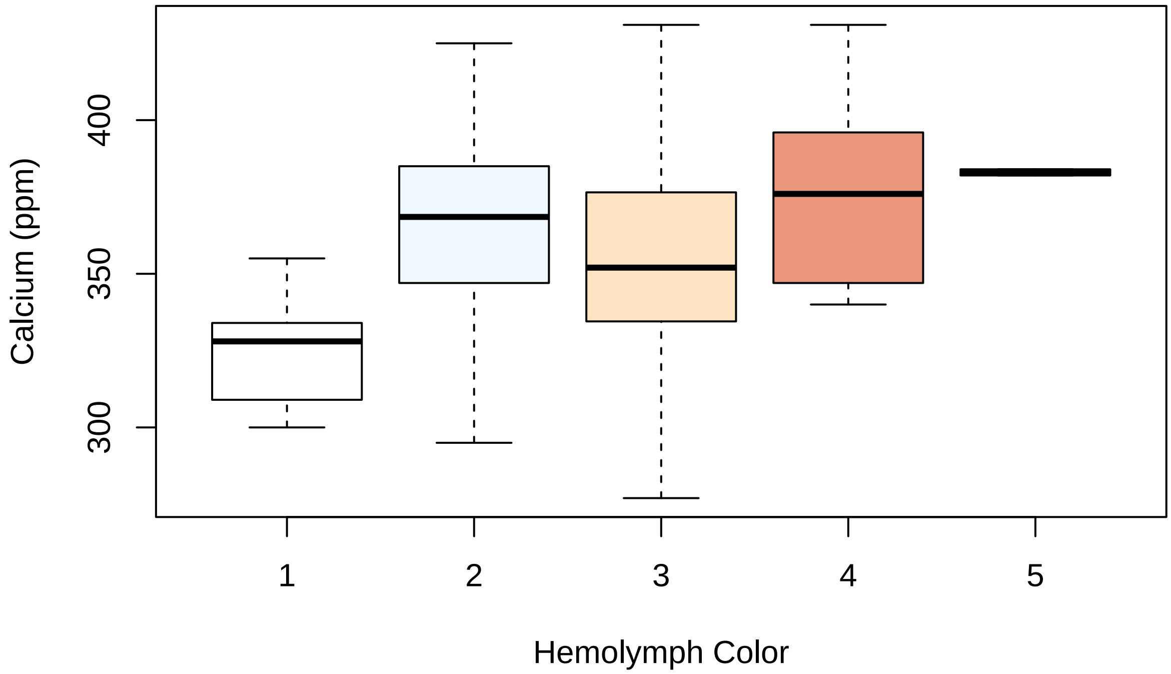
```
TukeyHSD(ri_col_aov_filt, conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = hemo_ri ~ fcol_filt, data = filt)
##
## $fcol_filt
##          diff        lwr      upr     p adj
## 2-1 0.6217576 -0.1444341 1.387949 0.1726677
## 3-1 1.6082925  0.9085657 2.308019 0.0000000
## 4-1 2.4842226  1.2713944 3.697051 0.0000004
## 5-1 3.1032702  1.1490905 5.057450 0.0001726
## 3-2 0.9865349  0.2918199 1.681250 0.0011217
## 4-2 1.8624650  0.6525213 3.072409 0.0003028
## 5-2 2.4815126  0.5291218 4.433903 0.0050213
## 4-3 0.8759301 -0.2930573 2.044917 0.2421847
## 5-3 1.4949777 -0.4322996 3.422255 0.2107252
## 5-4 0.6190476 -1.5479162 2.786011 0.9352081
```

Calcium plots

```
boxplot(ca ~ hemo_col, data = gc_dat_fin,
      xlab = "Hemolymph Color",
      ylab = "Calcium (ppm)",
      col= c('white', 'aliceblue', 'bisque', 'darksalmon','coral'),

      subset = period == "TRUE", outline = FALSE)
```



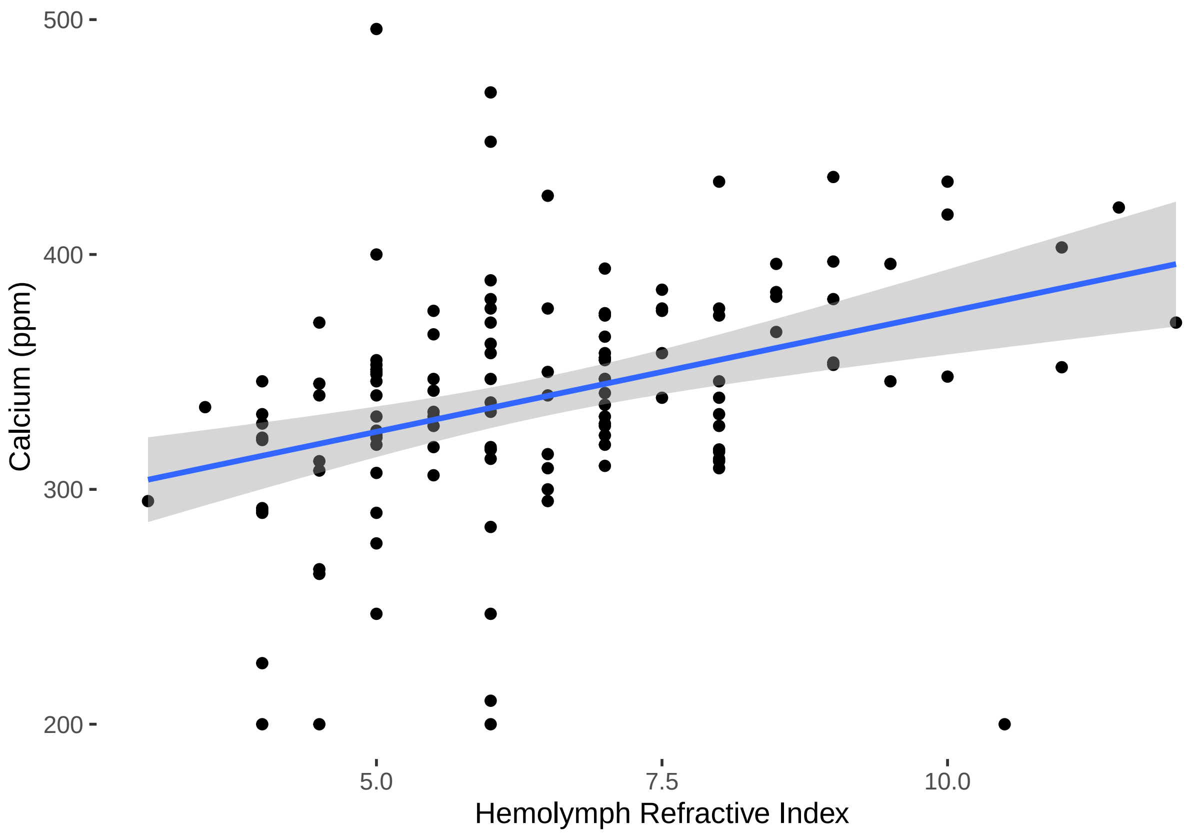
```
# appears that calcium increases as hemolymph color darkens
# n = 25 individual crabs
# n = 11, 10, 36, 6, 2
```

```
# Preliminary visualization from older dataset

gc_dat %>%
        # pipes molt data into ggplot
  ggplot(aes(x = RI, y = Ca)) +
        # populates plot with hemo ri and ca info
    geom_point() + geom_smooth(method=lm) +
        # specifies type of plot -- linear model with 95 percent confidence
        xlab("Hemolymph Refractive Index") +
        ylab ("Calcium (ppm)")  +
        # renames x and y labels and title
      xlim(3, 12) +
        # specifies x axis between 3 and 12 for more clear visualization
      theme(panel.grid.major = element_blank(),
              panel.grid.minor = element_blank(),
              panel.background = element_blank())
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 281 rows containing non-finite values (`stat_smooth()`).
## Warning: Removed 281 rows containing missing values ('geom_point()').
```



```
ca_ri_mod <- lm(Ca ~ RI, data = gc_dat)
nobs(ca_ri_mod) # tells that there are 127 observations used in this model
```

```
## [1] 127
```

```
summary(ca_ri_mod)
```

```
##
## Call:
## lm(formula = Ca ~ RI, data = gc_dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -180.608  -23.204    2.095   25.697  171.496
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  273.500     15.614  17.517  < 2e-16 ***
## RI            10.201      2.316   4.405 2.25e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 47.46 on 125 degrees of freedom
##   (281 observations deleted due to missingness)
## Multiple R-squared:  0.1344, Adjusted R-squared:  0.1274
## F-statistic:  19.4 on 1 and 125 DF,  p-value: 2.251e-05
```

```
anova(ca_ri_mod) # standard anova
```

```
## Analysis of Variance Table
##
## Response: Ca
##            Df Sum Sq Mean Sq F value    Pr(>F)
## RI          1  43702   43702  19.401 2.251e-05 ***
## Residuals 125 281571    2253
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
car::Anova(ca_ri_mod, type = 3) # type 3 anova for significance testing
```

```
## Anova Table (Type III tests)
##
## Response: Ca
##             Sum Sq  Df F value    Pr(>F)
## (Intercept) 691157   1 306.831 < 2.2e-16 ***
## RI           43702   1  19.401 2.251e-05 ***
## Residuals   281571 125
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# taken from all during molt cycle, leading to lots of variation
# pointing to trend; small r^2 due to assessing crabs at different times of
# molt cycle
```