P. mainlandi & Shrimp Data Analysis

R Markdown

Hypothesis #1: There is correlation between shrimp and goby size.

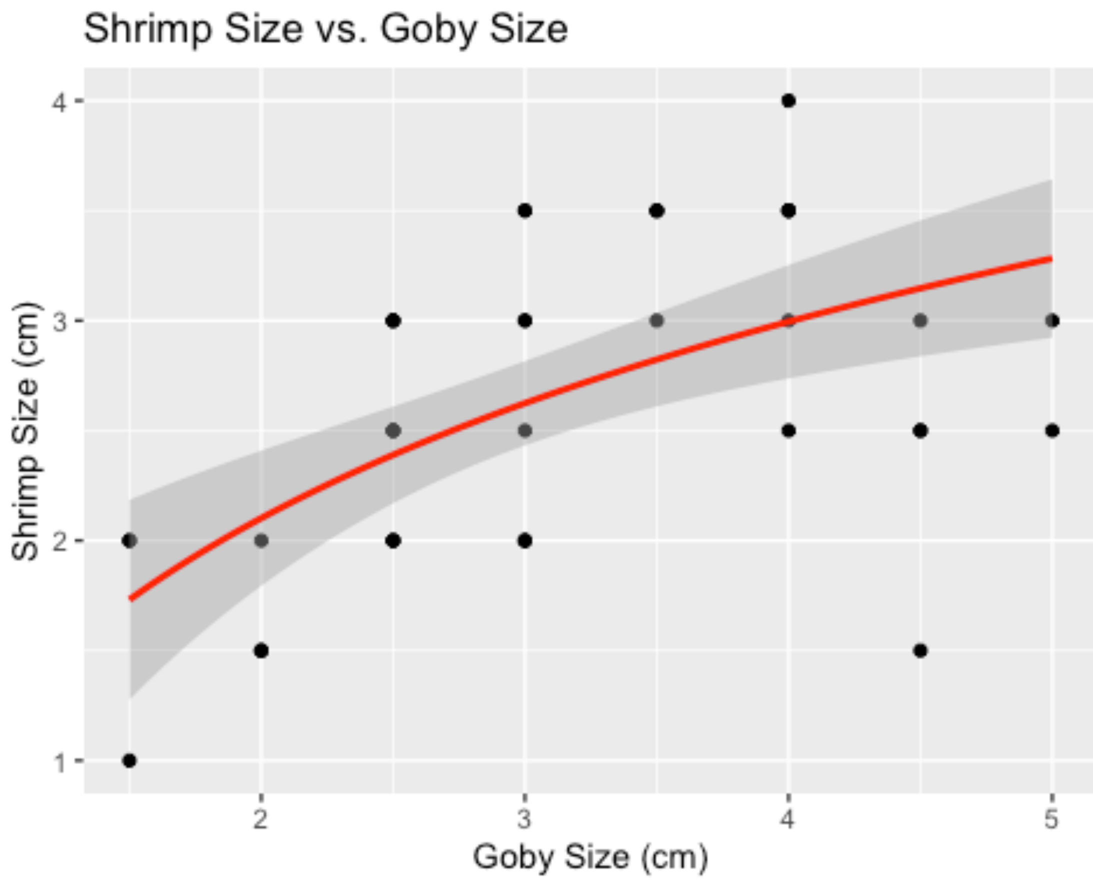
Step 1: Load data and prepare work space

```
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(plyr)

## ---------------------------------------------------------------------------
----

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first,
then dplyr:
## library(plyr); library(dplyr)

## ---------------------------------------------------------------------------
----

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

library(ggplot2)
library(readr)


data <- read_csv("mbio725_data.csv")

## Rows: 41 Columns: 4
```

```
## ── Column specification
────────────────────────────────────────────────────────
## Delimiter: ","
## dbl (4): goby_size_cm, shrimp_size_cm, avg_max_dist_cm, neighboring_gobies

##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.

View(data)
```

Step 2: Create logarithmic regression comparing shrimp vs. goby size

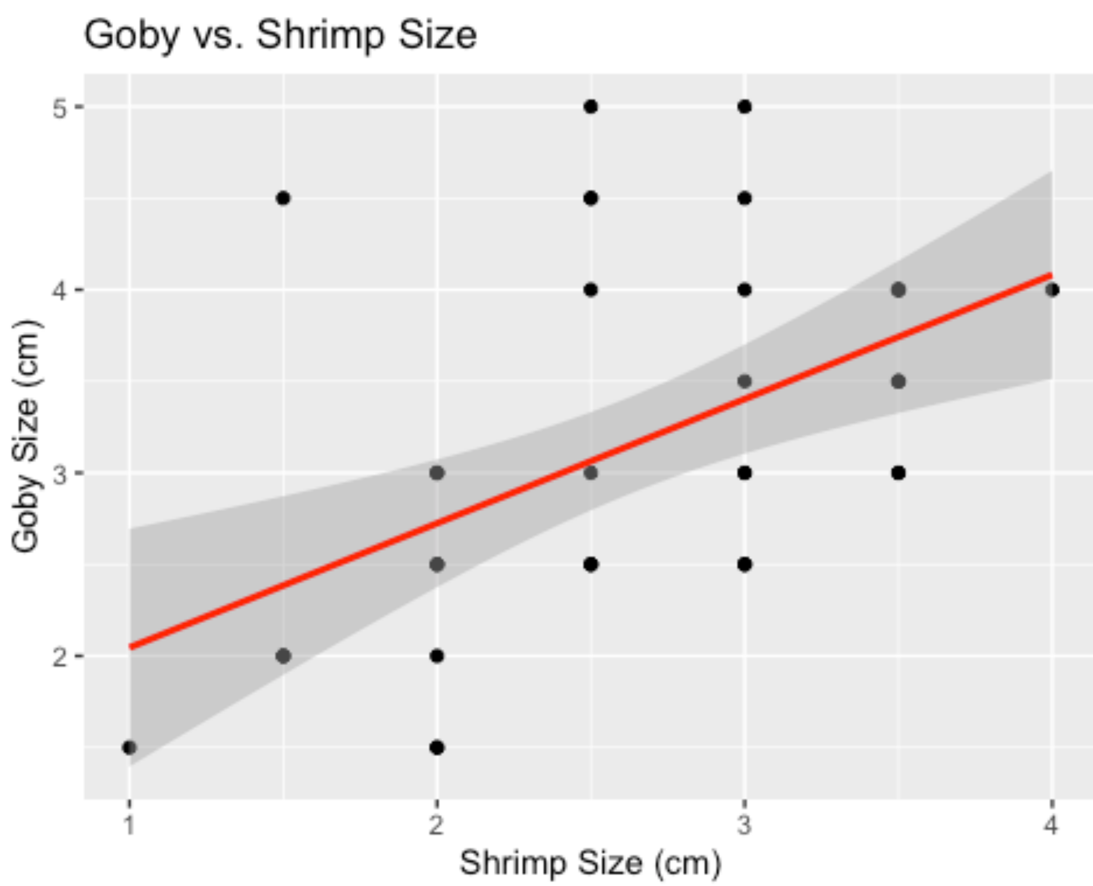
```
svg_log <- lm(shrimp_size_cm ~ log(goby_size_cm), data)
summary(svg_log)

##
## Call:
## lm(formula = shrimp_size_cm ~ log(goby_size_cm), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6465 -0.6017  0.1108  0.5052  1.0052
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.2085     0.3361   3.596 0.000897 ***
## log(goby_size_cm)   1.2885     0.2940   4.383 8.57e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6062 on 39 degrees of freedom
## Multiple R-squared:   0.33,  Adjusted R-squared:  0.3128
## F-statistic: 19.21 on 1 and 39 DF,  p-value: 8.572e-05

ggplot(data, aes(x = goby_size_cm, y = shrimp_size_cm)) +
  geom_point() +
  stat_smooth(method = "lm", formula=y~log(x), col = "red") +
  labs(title = "Shrimp Size vs. Goby Size",
       x = "Goby Size (cm)",
       y = "Shrimp Size (cm)")
```



Step 3: Create linear regression comparing goby vs. shrimp size

```
gvs <- lm(goby_size_cm ~ shrimp_size_cm, data)
summary(gvs)

##
## Call:
## lm(formula = goby_size_cm ~ shrimp_size_cm, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2242 -0.5636 -0.2242  0.2758  2.1152
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.3667     0.4917   2.780 0.008332 **
## shrimp_size_cm   0.6787     0.1808   3.754 0.000567 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8362 on 39 degrees of freedom
## Multiple R-squared:  0.2655, Adjusted R-squared:  0.2466
## F-statistic: 14.09 on 1 and 39 DF,  p-value: 0.0005667
```

```
ggplot(data, aes(x = shrimp_size_cm, y = goby_size_cm)) +
  geom_point() +
  stat_smooth(method = "lm", col = "red") +
  labs(title = "Goby vs. Shrimp Size",
       x = "Shrimp Size (cm)",
       y = "Goby Size (cm)")

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



Hypothesis #2: The maximum distance traveled from the burrow increases with increasing goby body size.

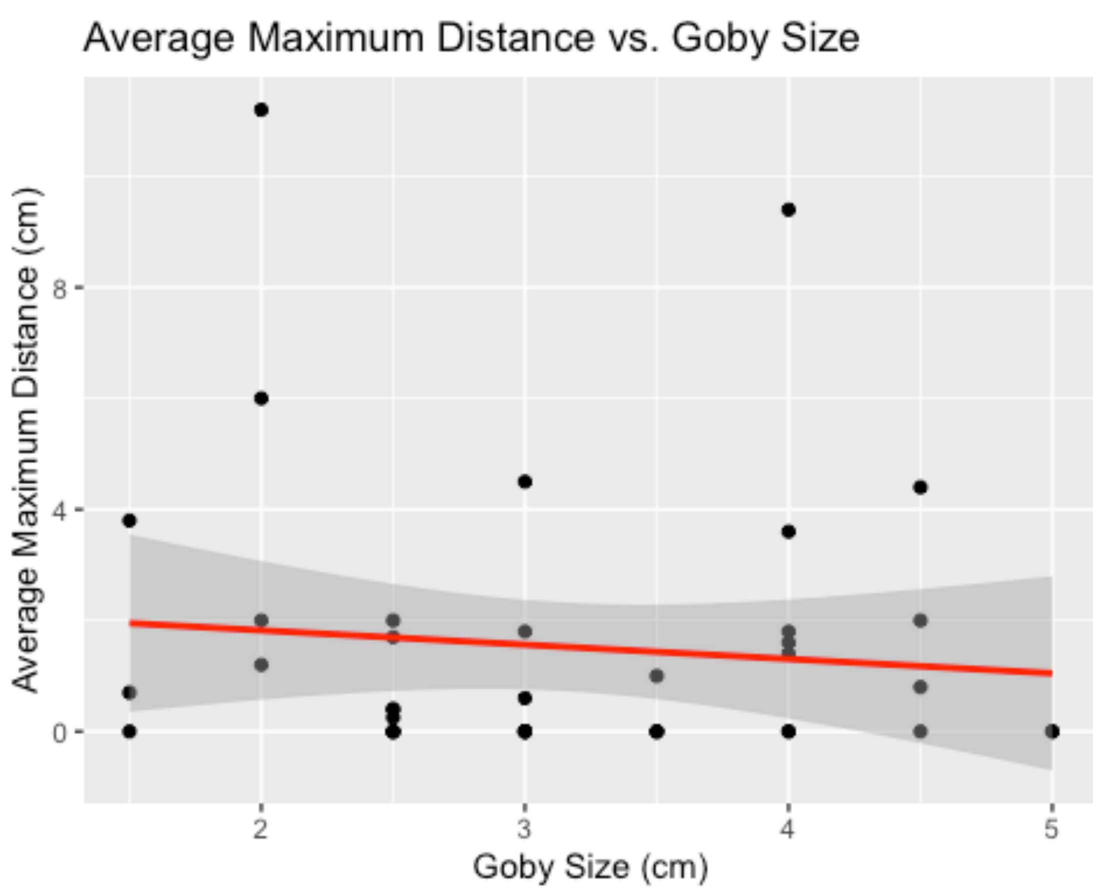
Step 4: Create linear regression comparing average maximum distance traveled vs. goby size

```
mdvg <- lm(avg_max_dist_cm ~ goby_size_cm, data)
summary(mdvg)

##
## Call:
## lm(formula = avg_max_dist_cm ~ goby_size_cm, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1.9504 -1.4424 -1.0473  0.2947  9.3786
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.3375     1.3635   1.714   0.0944 .
## goby_size_cm  -0.2580     0.4148  -0.622   0.5375
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.527 on 39 degrees of freedom
## Multiple R-squared:  0.009824,   Adjusted R-squared:  -0.01557
## F-statistic: 0.3869 on 1 and 39 DF,  p-value: 0.5375
```

```
ggplot(data, aes(x = goby_size_cm, y = avg_max_dist_cm)) +
  geom_point() +
  stat_smooth(method = "lm", col = "red") +
  labs(title = "Average Maximum Distance vs. Goby Size",
       x = "Goby Size (cm)",
       y = "Average Maximum Distance (cm)")
```

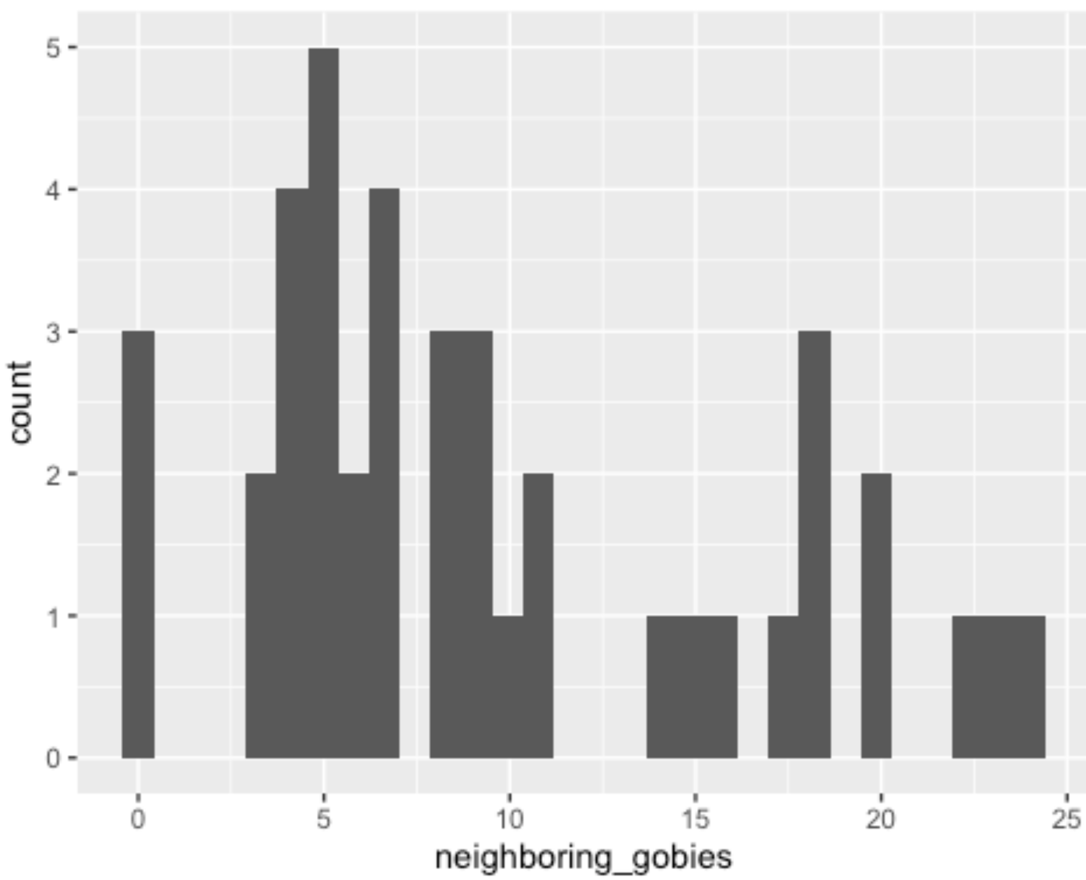
```
## `geom_smooth()` using formula 'y ~ x'
```



Hypothesis #3: The maximum distance traveled from the burrow increases in areas of higher goby density (within 1m^2)

Step 5: Create histogram to determine ideal grouping within density data

```
ggplot(data, aes(x = neighboring_gobies)) +
  geom_histogram()

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## testing out 0-5, 6-11, 12-24
## we must create new variable "density levels" (low, med, high) based on
neighboring_gobies variable

data_1 <- data %>%
  mutate(density = case_when(neighboring_gobies <= 5 ~ 'low',
                             neighboring_gobies <= 11 ~ 'med',
                             neighboring_gobies <= 24 ~ 'high'))
View(data_1)
```

Step 6: Run one-way ANOVA

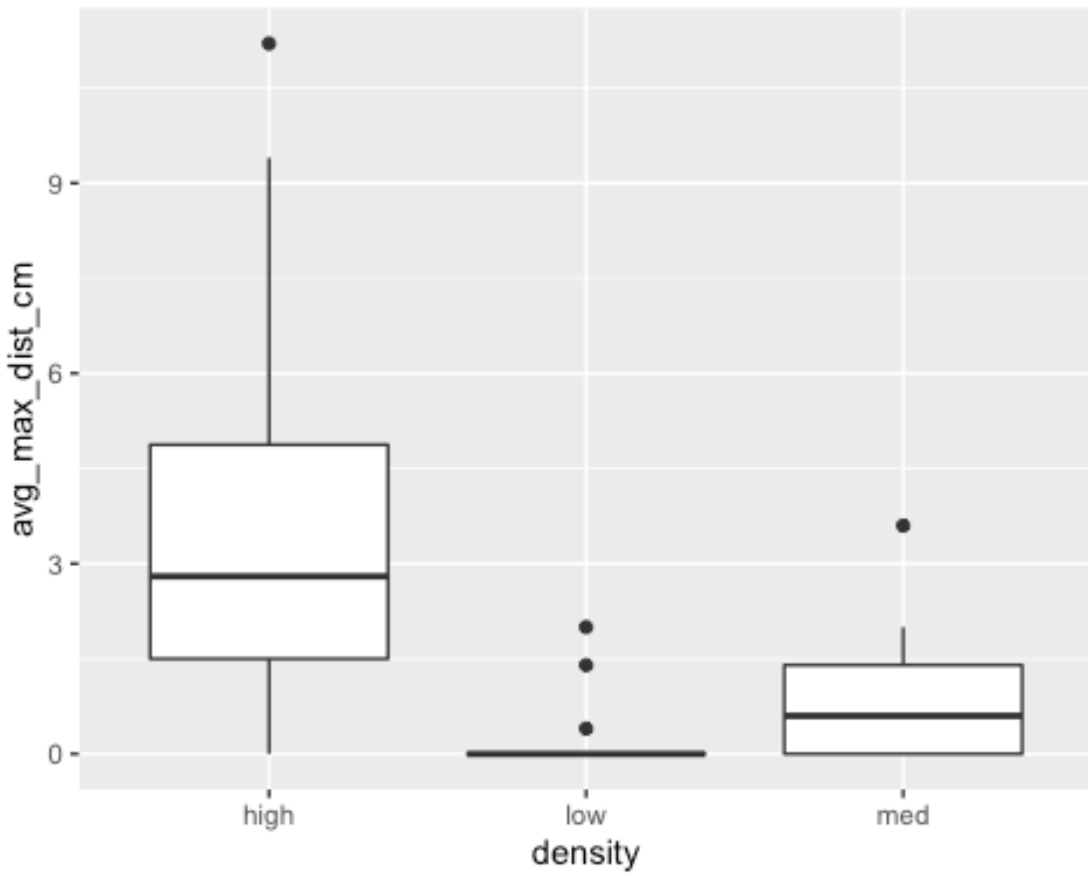
```
## Creating one-way ANOVA
one.way <- aov(avg_max_dist_cm ~ density, data = data_1)
summary(one.way)

##             Df Sum Sq Mean Sq F value   Pr(>F)    
## density      2  90.41   45.21   10.66 0.000212 ***
## Residuals   38 161.17    4.24                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## summary of ANOVA tells us there is a significant difference between the
three density levels for average maximum distance traveled
```

Step 7: Run post hoc test - Tukey HSD

```
## post hoc test to determine where the significant differences are
TukeyHSD(one.way)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## 
## Fit: aov(formula = avg_max_dist_cm ~ density, data = data_1)
## 
## $density
##                diff       lwr        upr     p adj
## low-high -3.5285714 -5.504481 -1.5526618 0.0002808
## med-high -2.9233333 -4.868607 -0.9780598 0.0021250
## med-low   0.6052381 -1.261246  2.4717226 0.7108764

## this tells us there is high significant difference in average maximum
distance traveled between low density and high density (p-value = 0.0002808),
there is also significant difference (but slightly less) in average maximum
distance traveled between medium density and high density (p-value =
0.0021250)
```

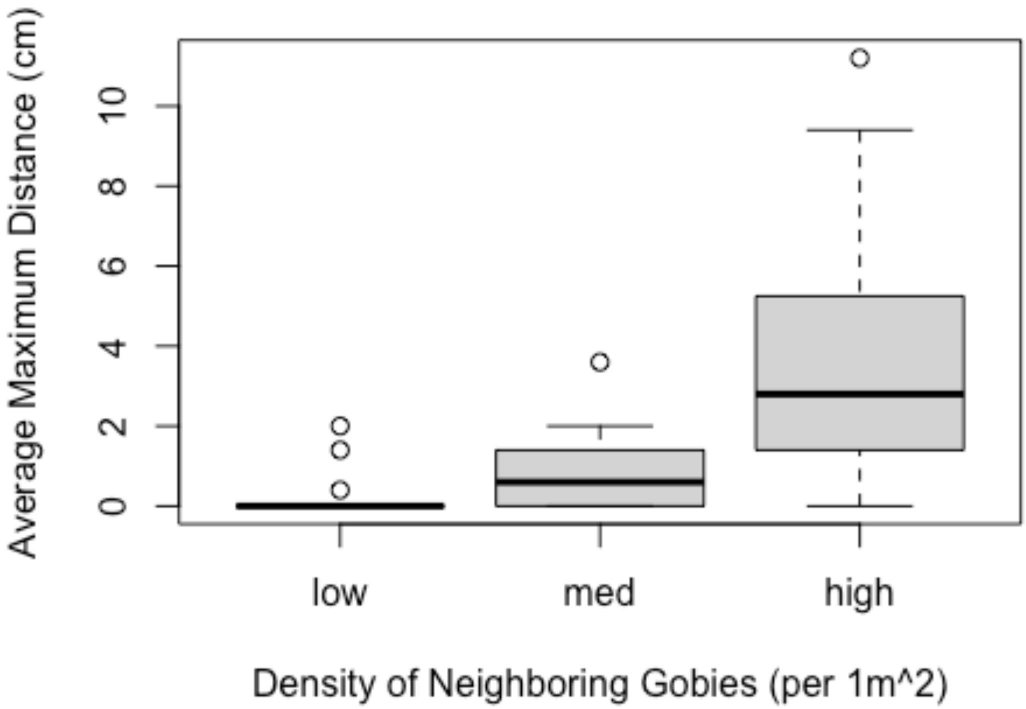
Step 8: Create box plots showing density data

```
ggplot(data_1, aes(x = density, y = avg_max_dist_cm)) +
  geom_boxplot()
```



Step 9: Reorder the groups from low to high & plot

```
#reorder low to high
data_1$density <- factor(data_1$density , levels=c("low", "med", "high"))

#The plot is now ordered !
boxplot(data_1$avg_max_dist_cm ~ data_1$density , ylab="Average Maximum
Distance (cm)" ,
    xlab="Density of Neighboring Gobies (per 1m^2)")
```



Average Maximum Distance (cm)
0
2
4
6
8
10
low
med
high
Density of Neighboring Gobies (per 1m^2)