Week 2: Data Cleaning & Exploration

2022-08-01

# Step 0: Clean/inspect data

Sea\_urchins <- read.csv("Desktop/Design & Analyses/Datasets/Sea\_urchins.csv")  
  
summary(Sea\_urchins)

## Site Area Size Code   
## Length:1744 Length:1744 Length:1744 Length:1744   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## Month TD TW GW   
## Min. : 1.000 Min. :30.00 Length:1744 Min. : 0.010   
## 1st Qu.: 3.000 1st Qu.:40.00 Class :character 1st Qu.: 0.790   
## Median : 6.000 Median :50.00 Mode :character Median : 1.715   
## Mean : 6.081 Mean :48.69 Mean : 2.292   
## 3rd Qu.: 9.000 3rd Qu.:56.00 3rd Qu.: 3.360   
## Max. :12.000 Max. :74.00 Max. :13.100   
## NA's :320   
## GSI Sex   
## Min. : 0.030 Length:1744   
## 1st Qu.: 1.597 Class :character   
## Median : 2.810 Mode :character   
## Mean : 3.357   
## 3rd Qu.: 4.635   
## Max. :11.860   
## NA's :320

# Check Datatypes   
  
result="hide"  
fig.show ="hide"

# Converting into factors   
Sea\_urchins$Site <-factor(Sea\_urchins$Site)  
Sea\_urchins$Area <-factor(Sea\_urchins$Area)  
Sea\_urchins$Month <-factor(Sea\_urchins$Month)  
Sea\_urchins$TD <- factor(Sea\_urchins$TD)  
  
result="hide"  
fig.show ="hide"

# Collapsing  
  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4  
## ✔ tibble 3.1.8 ✔ dplyr 1.0.9  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.0  
## ✔ readr 2.1.2 ✔ forcats 0.5.1  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

Sea\_urchins$Area <- fct\_collapse(Sea\_urchins$Area, High = c("A" , "B"), Low = c("C","D"))  
summary(Sea\_urchins$Area)

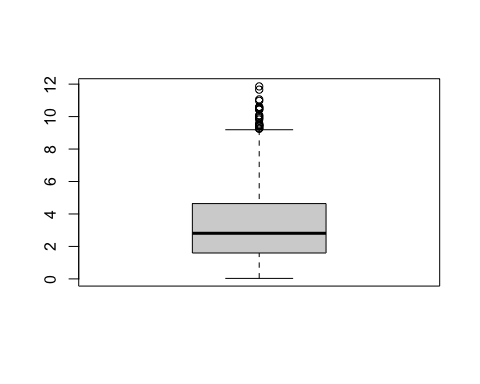
## High Low   
## 875 869

# Areas with high and low fishing pressure  
  
result="hide"  
fig.show ="hide"

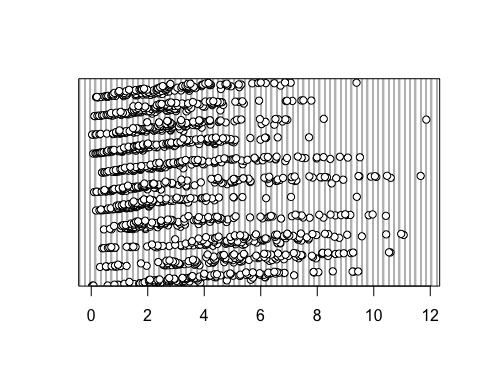
Step 0 Conclusion: summarized the data, converted to factors, collapsed the data

# Step 1: Finding the Outliers

# Finding the outliers for GSI boxplot and dotchart  
  
x <-Sea\_urchins$GSI  
boxplot(x)



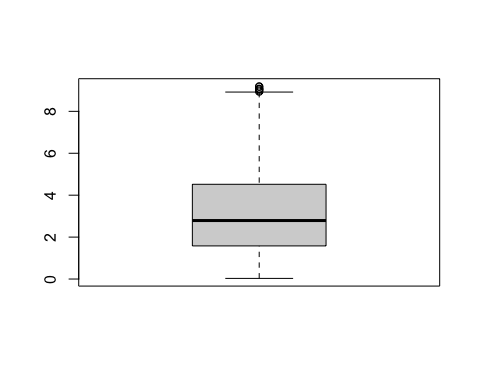
dotchart(x)



result="hide"  
fig.show ="hide"

# Sensitivity Analysis  
  
library(tidyverse)  
Q3 <- quantile(Sea\_urchins$GSI, probs = .75, na.rm = TRUE)  
IQR <- IQR(Sea\_urchins$GSI, na.rm = TRUE)  
seaurchins1 <- subset(Sea\_urchins$GSI, Sea\_urchins$GSI < (Q3 + 1.5\*IQR))  
   
result="hide"  
fig.show ="hide"

boxplot(seaurchins1)



# Outliers have been removed  
  
result="hide"  
fig.show ="hide"

Step 1 Conclusion: made a boxplot and dotchart, found outliers, and removed outliers

# Step 2: Homogeneity of variance

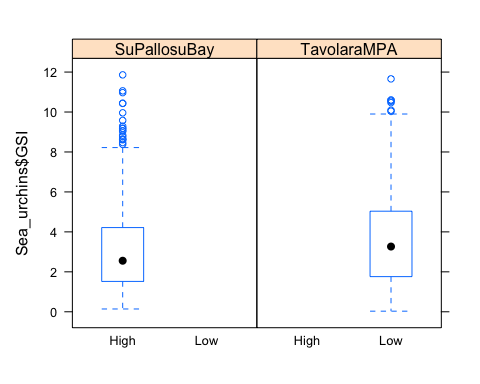
library(tidyverse)  
Sea\_urchins$Area <- fct\_collapse(Sea\_urchins$Area, High = c("A" , "B"), Low = c("C","D"))

## Warning: Unknown levels in `f`: A, B, C, D

summary(Sea\_urchins$Area)

## High Low   
## 875 869

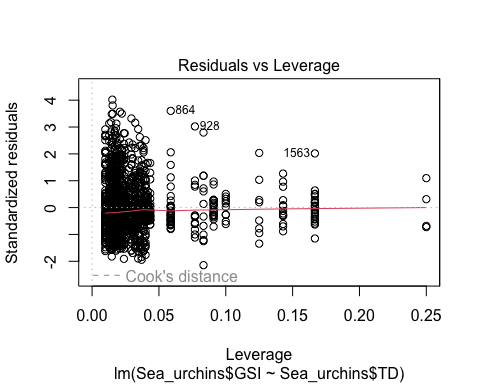
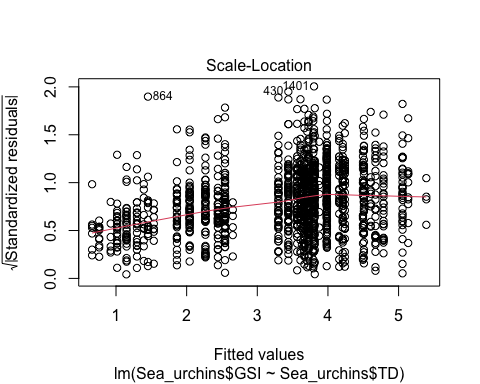
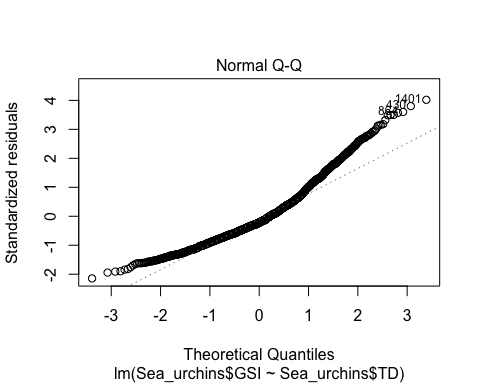
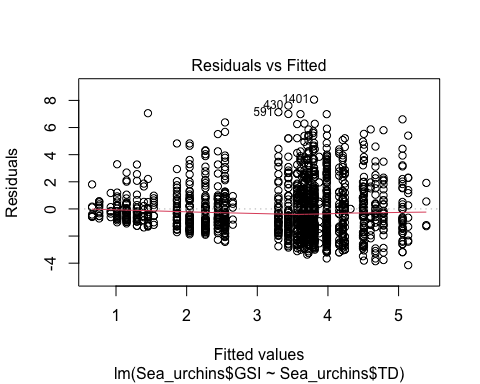
library(lattice)  
bwplot(Sea\_urchins$GSI ~ Sea\_urchins$Area | Sea\_urchins$Site)



result="hide"  
fig.show ="hide"

# Creating a residual plot vs fitted plot for TD  
model <- lm(Sea\_urchins$GSI ~ Sea\_urchins$TD, data = Sea\_urchins)  
  
plot(model)

## Warning: not plotting observations with leverage one:  
## 627, 1290

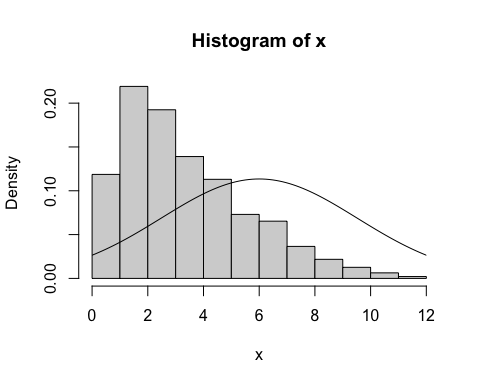


result="hide"  
fig.show ="hide"

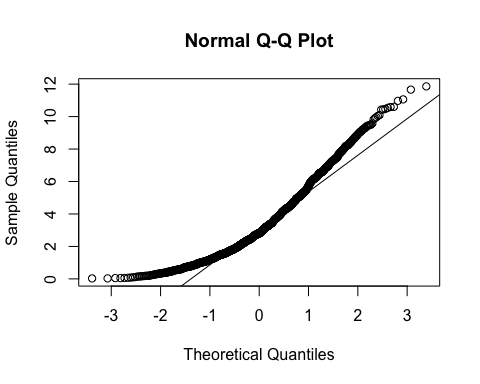
Step 2 Conclusion: we have equal variances within each group

# Step 3: Checking for Normality

# Checking for normality   
x <-Sea\_urchins$GSI  
  
{hist(x, freq=FALSE)   
 curve(dnorm(x,mean=mean(x), sd=sd(x)),   
 add=TRUE, type = "l")}



# QQ Normality Plot  
qqnorm(x)  
qqline(x)



result="hide"  
fig.show ="hide"

Step 3 Conclusion: data is not normally distributed

# Step 4: Are there lots of zeros in the data?

Does not apply because we are not using count/discrete data, we’re using continuous data

# Step 5: Is there colinearity among the covariates?

library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

## The following object is masked from 'package:purrr':  
##   
## some

model <- lm(Sea\_urchins$GSI ~ Sea\_urchins$GW + Sea\_urchins$TD + Sea\_urchins$TW, data = Sea\_urchins)  
# vif(model) --> returned an error "aliased coefficients in the model" so had to remove TW  
  
# VIF of TW is highest (14.2) so removing that variable  
  
model <- lm(Sea\_urchins$GSI ~ Sea\_urchins$GW + Sea\_urchins$TD, data = Sea\_urchins)  
vif(model)

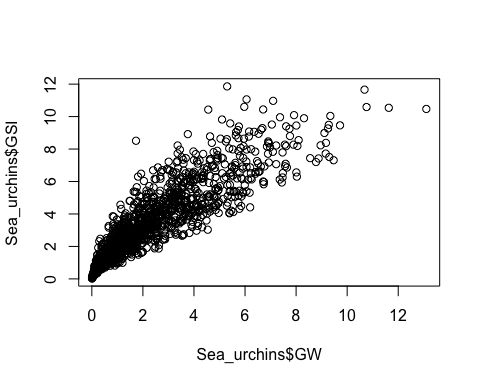
## GVIF Df GVIF^(1/(2\*Df))  
## Sea\_urchins$GW 1.948366 1 1.395839  
## Sea\_urchins$TD 1.948366 44 1.007608

# VIF of GW and TD are now both less than 3 so this step is complete  
  
result="hide"  
fig.show ="hide"

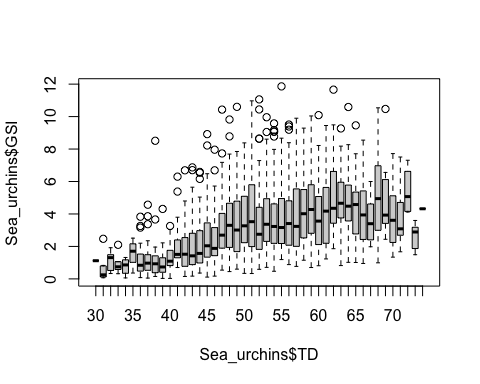
Step 5 Conclusion: had to drop TW to reduce collinearity among the covariates

# Step 6: What are the relationships between x and y variables?

# Plot response vs. each covariate  
  
plot(Sea\_urchins$GSI ~ Sea\_urchins$GW)



plot(Sea\_urchins$GSI ~ Sea\_urchins$TD)

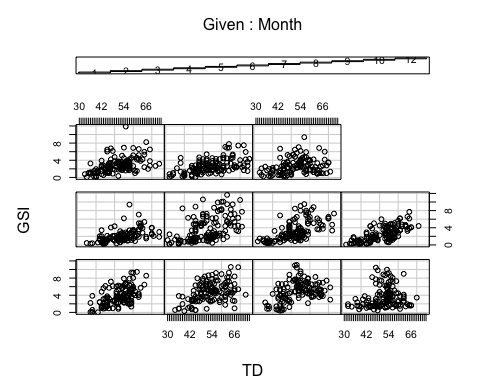


result="hide"  
fig.show ="hide"

Step 6 Conclusions: GW and GSI show a positive non-linear relationship, as do TD and GSI

# Step 7: Should we consider interactions?

# Coplot  
  
coplot(GSI ~ TD | Month, data = Sea\_urchins)



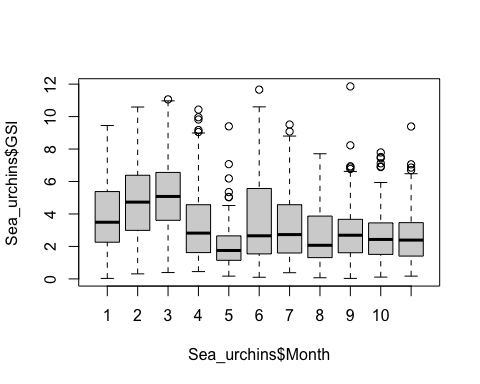
##   
## Missing rows: 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 285, 286, 287, 288, 289, 290, 291, 292, 293, 294, 295, 296, 297, 298, 299, 300, 301, 302, 303, 304, 305, 306, 307, 308, 309, 310, 311, 312, 313, 314, 315, 316, 317, 443, 444, 445, 446, 447, 448, 449, 450, 451, 452, 453, 454, 455, 456, 457, 458, 459, 460, 461, 462, 463, 464, 465, 466, 467, 468, 469, 470, 471, 472, 473, 474, 475, 476, 605, 606, 607, 608, 609, 610, 611, 612, 613, 614, 615, 616, 617, 618, 619, 620, 621, 622, 623, 624, 625, 626, 627, 628, 629, 630, 631, 632, 633, 634, 635, 752, 753, 754, 755, 756, 757, 758, 759, 760, 761, 762, 763, 764, 765, 766, 767, 768, 769, 770, 771, 772, 773, 774, 775, 776, 777, 778, 779, 780, 781, 782, 783, 784, 785, 786, 787, 788, 789, 790, 929, 930, 931, 932, 933, 934, 935, 936, 937, 938, 939, 940, 941, 942, 943, 944, 945, 946, 947, 948, 949, 950, 951, 952, 953, 954, 1091, 1092, 1093, 1094, 1095, 1096, 1097, 1098, 1099, 1100, 1101, 1102, 1103, 1104, 1105, 1106, 1107, 1108, 1109, 1110, 1111, 1112, 1113, 1114, 1253, 1254, 1255, 1256, 1257, 1258, 1259, 1260, 1261, 1262, 1263, 1264, 1265, 1266, 1267, 1268, 1269, 1270, 1271, 1272, 1273, 1408, 1409, 1410, 1411, 1412, 1413, 1414, 1415, 1416, 1417, 1418, 1419, 1420, 1421, 1422, 1423, 1424, 1425, 1426, 1427, 1428, 1429, 1564, 1565, 1566, 1567, 1568, 1569, 1570, 1571, 1572, 1573, 1574, 1575, 1576, 1577, 1578, 1579, 1580, 1581, 1582, 1583, 1584, 1585, 1586, 1587, 1715, 1716, 1717, 1718, 1719, 1720, 1721, 1722, 1723, 1724, 1725, 1726, 1727, 1728, 1729, 1730, 1731, 1732, 1733, 1734, 1735, 1736, 1737, 1738, 1739, 1740, 1741, 1742, 1743, 1744

result="hide"  
fig.show ="hide"

Step 7 Conclusion: interactions to be discussed in week 4

# Step 8: Are observations of the response variable independent?

plot(Sea\_urchins$GSI ~ Sea\_urchins$Month)  
  
boxplot(Sea\_urchins$GSI ~ Sea\_urchins$Month)



result="hide"  
fig.show ="hide"

Step 8 Conclusions: measures do not appear to be repeated, observations appear independent

# Would an ANOVA be an appropriate test to examine the research question?

* Unsure because we have a large sample size, but the data isn’t normally distributed