Data Vitualisation and Analysis

Felix Ehis Osawaru

06/04/2022

```
knitr::opts_chunk$set(echo = TRUE)

library(ggplot2)

library(plyr)

## Warning: package 'plyr' was built under R version 4.1.3

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.1.3

##
## Attaching package: 'dplyr'

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

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

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

library(lattice)

library(Rmisc)

## Warning: package 'Rmisc' was built under R version 4.1.3

plankton <- read.csv("plankton.csv", header=T, stringsAsFactors = T)

                                QUESTION 1

summary(plankton)

##      Sample      Pseudonitzschia.A.Sp Alexandrium.Sp      Robgordia.Sp
##  Min.   :  1.0   Min.   :  787.3      Min.   :    0.00   Min.   :   7.3
##  1st Qu.:189.2   1st Qu.: 1350.9      1st Qu.:    0.00   1st Qu.: 185.7
##  Median :377.5   Median : 2180.2      Median :    0.00   Median : 276.6
##  Mean   :377.5   Mean   : 3761.5      Mean   :  145.96   Mean   : 425.0
```

```
##  3rd Qu.:565.8   3rd Qu.: 4206.9      3rd Qu.:   40.04   3rd Qu.: 458.6
##  Max.   :754.0   Max.   :35056.5      Max.   :30530.50   Max.   :3611.7
##
##    Water.Temp               Species        Region         Site
##  Min.   :-0.50   Common cockles : 16   SIC    :498   SI-288 : 57
##  1st Qu.: 9.70   Common mussels :656   AGB    : 67   SI-327 : 53
##  Median :12.10   Pacific oysters: 76   CESLH  : 66   SI-242 : 51
##  Mean   :12.17   Razors         :  6   HCSL   : 43   SI-035 : 47
##  3rd Qu.:14.90                        HCS    : 33   SI-080 : 36
##  Max.   :24.60                        HCL    : 26   SI-326 : 36
##                                       (Other): 21   (Other):474
##       day            month             year               period
##  Min.   : 1.00   Min.   : 3.000   Min.   :2009   1st half year:250
##  1st Qu.: 9.00   1st Qu.: 6.000   1st Qu.:2011   2nd half year:504
##  Median :16.00   Median : 7.000   Median :2015
##  Mean   :16.35   Mean   : 7.042   Mean   :2015
##  3rd Qu.:23.00   3rd Qu.: 8.000   3rd Qu.:2020
##  Max.   :31.00   Max.   :10.000   Max.   :2021
##

str(plankton)

## 'data.frame':    754 obs. of  12 variables:
##  $ Sample              : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Pseudonitzschia.A.Sp: num  2482 2128 8906 1407 2551 ...
##  $ Alexandrium.Sp      : num  0 120.1 60.1 20 20 ...
##  $ Robgordia.Sp        : num  368 218 2408 178 375 ...
##  $ Water.Temp          : num  20.4 8.6 14.9 17.6 13.7 20.2 14.5 14.3 12.4
17.8 ...
##  $ Species             : Factor w/ 4 levels "Common cockles",..: 2 2 2 2 2
2 2 2 2 2 ...
##  $ Region              : Factor w/ 11 levels "AGB","CESLH",..: 7 8 8 8 2
11 2 8 2 8 ...
##  $ Site                : Factor w/ 59 levels "AB-029","AB-041",..: 18 57
57 57 22 41 22 57 22 57 ...
##  $ day                 : int  22 27 5 11 26 26 1 1 8 8 ...
##  $ month               : int  4 4 5 5 5 5 6 6 6 6 ...
##  $ year                : int  2009 2009 2009 2009 2009 2009 2009 2009 2009
2009 ...
##  $ period              : Factor w/ 2 levels "1st half year",..: 1 1 1 1 1
1 1 1 1 1 ...

class(plankton)

## [1] "data.frame"
```

##Pseudonitzschia.A.Sp

```
summary(plankton$Pseudonitzschia.A.Sp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   787.3  1350.9  2180.2  3761.5  4206.9 35056.5
```

Obtaining the statistical value for plankton and Pseudonitzschia.A.Sp

```
mean(plankton$Pseudonitzschia.A.Sp)

## [1] 3761.542

median(plankton$Pseudonitzschia.A.Sp)

## [1] 2180.25

sd(plankton$Pseudonitzschia.A.Sp)

## [1] 4391.953

max(plankton$Pseudonitzschia.A.Sp)

## [1] 35056.5

IQR(plankton$Pseudonitzschia.A.Sp)

## [1] 2856.025

range(plankton$Pseudonitzschia.A.Sp)

## [1]   787.3 35056.5

var(plankton$Pseudonitzschia.A.Sp)

## [1] 19289250
```

##Alexandrium.Sp

```
summary(plankton$Alexandrium.Sp)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     0.00     0.00     0.00   145.96    40.04 30530.50
```

Obtaining the statistical value for plankton and Alexandrium.Sp

```
mean(plankton$Alexandrium.Sp)

## [1] 145.9548

median(plankton$Alexandrium.Sp)

## [1] 0

sd(plankton$Alexandrium.Sp)

## [1] 1204.903

min(plankton$Alexandrium.Sp)

## [1] 0
```

```
max(plankton$Alexandrium.Sp)

## [1] 30530.5

IQR(plankton$Alexandrium.Sp)

## [1] 40.04

range(plankton$Alexandrium.Sp)

## [1]     0.0 30530.5

var(plankton$Alexandrium.Sp)

## [1] 1451791
```

Robgordia.Sp

```
summary(plankton$Robgordia.Sp)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     7.3   185.7   276.6   425.0   458.6  3611.7
```

Obtaining the statistical value for plankton and Robgordia.Sp

```
mean(plankton$Robgordia.Sp)

## [1] 424.9891

median(plankton$Robgordia.Sp)

## [1] 276.65

sd(plankton$Robgordia.Sp)

## [1] 450.8656

min(plankton$Robgordia.Sp)

## [1] 7.3

max(plankton$Robgordia.Sp)

## [1] 3611.7

IQR(plankton$Robgordia.Sp)

## [1] 272.925

range(plankton$Robgordia.Sp)

## [1]    7.3 3611.7

var(plankton$Robgordia.Sp)

## [1] 203279.8
```

water.Temp

```
summary(plankton$Water.Temp)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -0.50    9.70   12.10   12.17   14.90   24.60

mean(plankton$Water.Temp)

## [1] 12.17361

median(plankton$Water.Temp)

## [1] 12.1

sd(plankton$Water.Temp)

## [1] 4.122823

min(plankton$Water.Temp)

## [1] -0.5

max(plankton$Water.Temp)

## [1] 24.6

IQR(plankton$Water.Temp)

## [1] 5.2

range(plankton$Water.Temp)

## [1] -0.5 24.6

var(plankton$Water.Temp)

## [1] 16.99767
```

Region

```
counts <- table(plankton$Region)
counts

##
##   AGB CESLH CESUB    FC   HCL  HCRC   HCS  HCSL   NAC   SAC   SIC
##    67    66     2     2    26     8    33    43     5     4   498

prop.table(counts)

##
##         AGB       CESLH       CESUB          FC         HCL        HCRC
HCS
## 0.08885942 0.08753316 0.00265252 0.00265252 0.03448276 0.01061008
0.04376658
```

```
##       HCSL        NAC        SAC        SIC
## 0.05702918 0.00663130 0.00530504 0.66047745
```

##Species

```
counts <- table(plankton$Species)
counts

##
##  Common cockles  Common mussels Pacific oysters          Razors
##              16             656              76               6

prop.table(counts)

##
##  Common cockles  Common mussels Pacific oysters          Razors
##      0.02122016      0.87002653      0.10079576      0.00795756
```

Year

```
counts <- table(plankton$year)
counts

##
## 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021
##   63   63   94   61   22   36   48   23   87   33   28   90  106

prop.table(counts)

##
##       2009       2010       2011       2012       2013       2014
2015
## 0.08355438 0.08355438 0.12466844 0.08090186 0.02917772 0.04774536
0.06366048
##       2016       2017       2018       2019       2020       2021
## 0.03050398 0.11538462 0.04376658 0.03713528 0.11936340 0.14058355
```

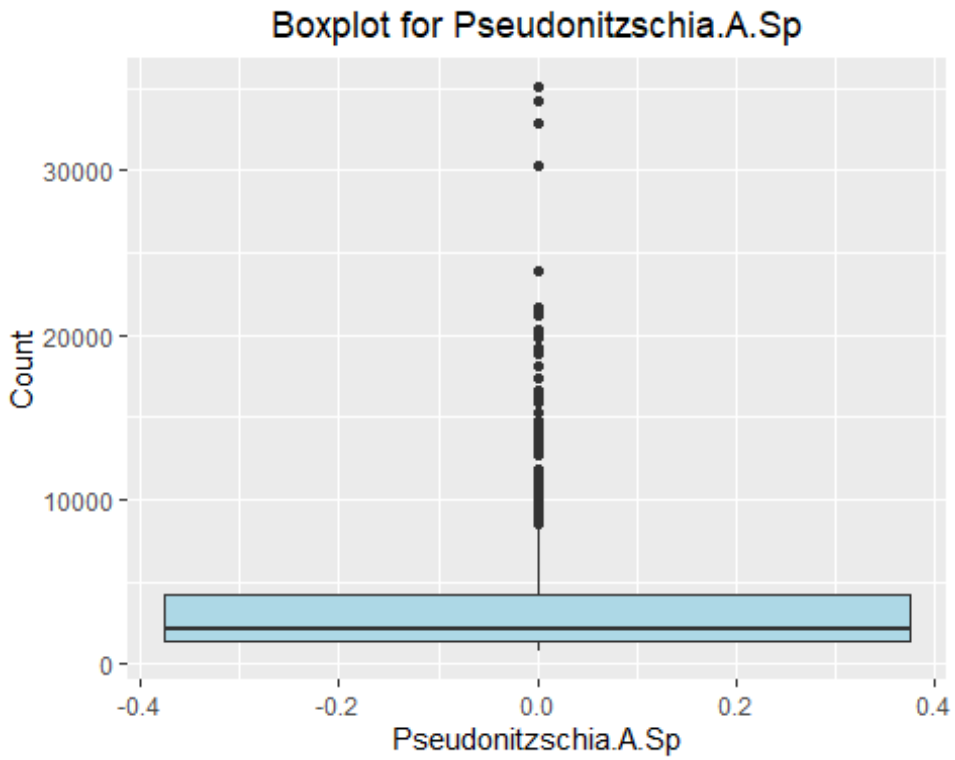
Period

```
counts <- table(plankton$period)
counts

##
## 1st half year 2nd half year
##           250           504

prop.table(counts)

##
## 1st half year 2nd half year
##      0.331565      0.668435
```

QUESTION 2

Using Boxplot to show the distribution of Pseudonitzschia

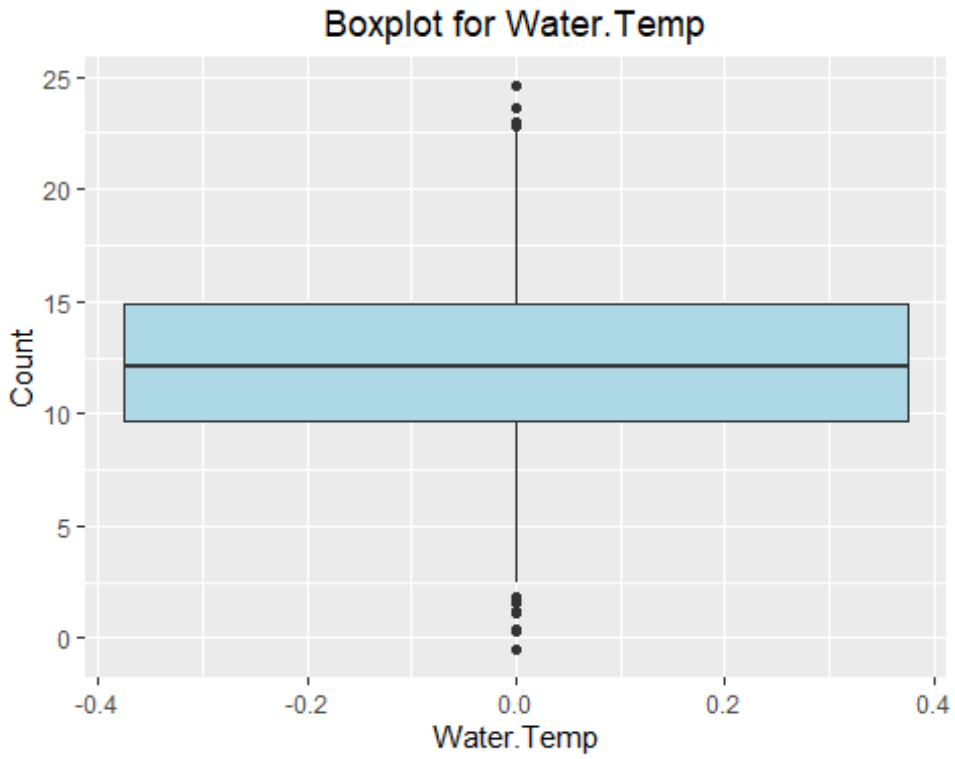
```
p <- ggplot(plankton, aes(y=Pseudonitzschia.A.Sp))
p <- p + geom_boxplot(fill="lightblue",)
p <- p + labs(x="Pseudonitzschia.A.Sp",
         y="Count",
         title = "Boxplot for Pseudonitzschia.A.Sp")+
         theme(plot.title = element_text(hjust = 0.5))
p
```



The distribution is positively skewed, the number of outliers is within the upper bound of the data, and the median does not divide the box evenly. That is, the mean, the standard deviation is higher than the median.

Water.Temp

```
p <- ggplot(plankton, aes(y=Water.Temp))
p <- p + geom_boxplot(fill="lightblue",)
p <- p + labs(x="Water.Temp",
         y="Count",
         title = "Boxplot for Water.Temp")+
         theme(plot.title = element_text(hjust = 0.5))
p
```



The distribution is symmetric because the whiskers and outliers on the left are almost the same as on the right, and there is an equal amount of data in each quadrant.

QUESTION 3

Using Univeriate statistics to compare Pseudonitzchia data for 2021 and before 2021

```
Pseudonitzchia.A.sp2021 <- plankton$Pseudonitzschia.A.Sp[plankton$year==
2021]
summary(Pseudonitzchia.A.sp2021)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   816.5  1494.2  2556.6  4883.7  6379.9 21709.6
```

##Obtaining the statistical value for plankton$Pseudonitzschia.A.Sp for 2021

```
sd(Pseudonitzchia.A.sp2021)

## [1] 5094.941

range(Pseudonitzchia.A.sp2021)

## [1]   816.5 21709.6

var(Pseudonitzchia.A.sp2021)

## [1] 25958427
```

```
IQR(Pseudonitzchia.A.sp2021)

## [1] 4885.7
```

Pseudonitzchia distribution Before 2021

```
Pseudonitzchia.A.sp2021 <- plankton$Pseudonitzschia.A.Sp[plankton$year !=
2021]
summary(Pseudonitzchia.A.sp2021)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   787.3  1344.7  2129.4  3578.0  3898.2 35056.5
```

##Obtaining the statistical value for plankton$Pseudonitzschia.A.Sp before 2021

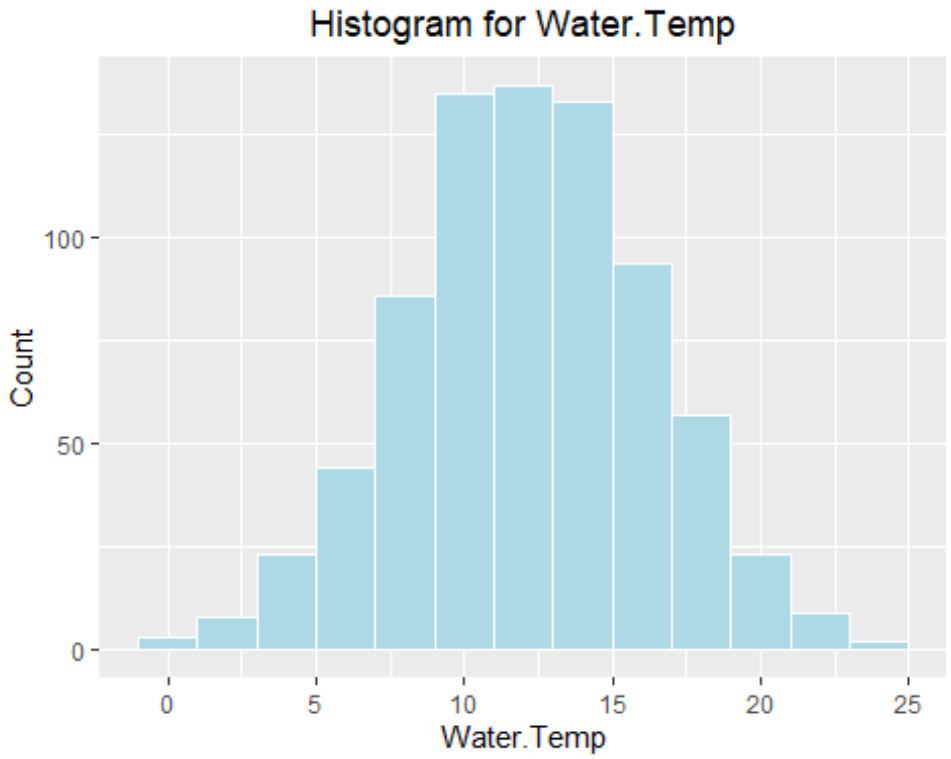
```
sd(Pseudonitzchia.A.sp2021)

## [1] 4242.249

range(Pseudonitzchia.A.sp2021)

## [1]   787.3 35056.5

var(Pseudonitzchia.A.sp2021)

## [1] 17996673

IQR(Pseudonitzchia.A.sp2021)

## [1] 2553.5
```

Comparing the data from 2021 to those before 2021, it can be seen that the mean, median, and standard deviation of 2021 are higher than the data of previous years. It suggests that Pseudonitzschia.A.Sp species in 2021 are more dispersed than in previous years.

```
                                QUESTION 4
```

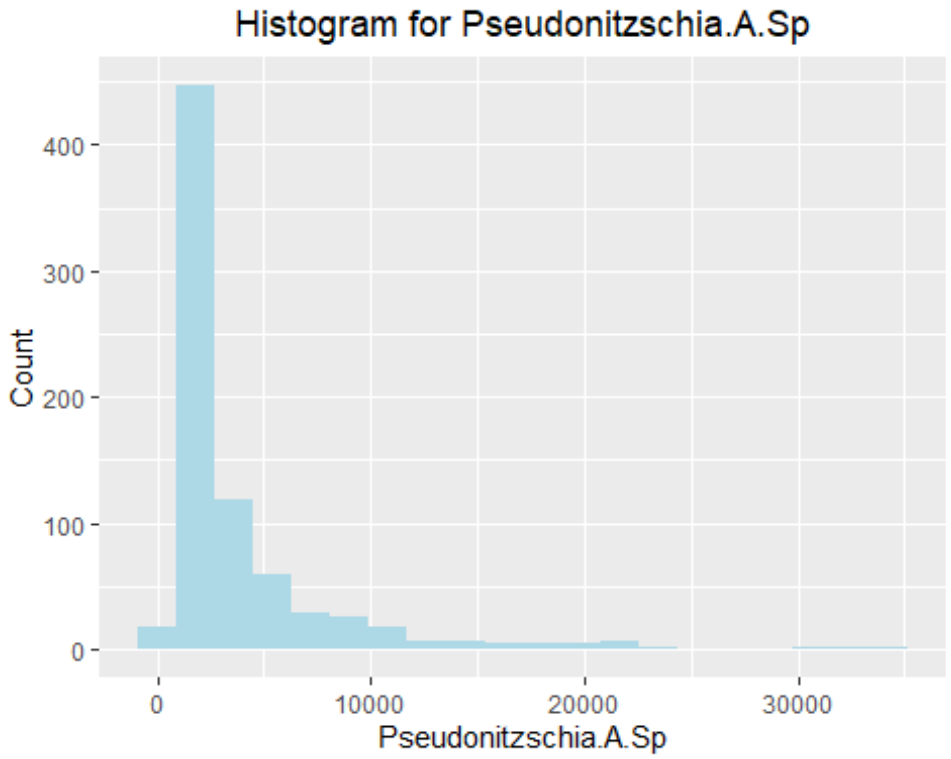
Plotting the Histogram of Normal and Skewed distribution

```
p <- ggplot(data =plankton, aes(Water.Temp))
p <- p + geom_histogram(colour="white", fill="lightblue", binwidth =2)
p <- p + labs(x="Water.Temp",
         y="Count",
         title = "Histogram for Water.Temp")+
         theme(plot.title = element_text(hjust = 0.5))
p
```



This is a normal distribution because its bell curve has a peak, and the mean, median, and mode are equal.

```
p <- ggplot(data =plankton, aes(Pseudonitzschia.A.Sp))
p <- p + geom_histogram( fill="lightblue",bins=20)
p <- p + labs(x="Pseudonitzschia.A.Sp",
         y="Count",
         title = "Histogram for Pseudonitzschia.A.Sp")+
         theme(plot.title = element_text(hjust = 0.5))
p
```

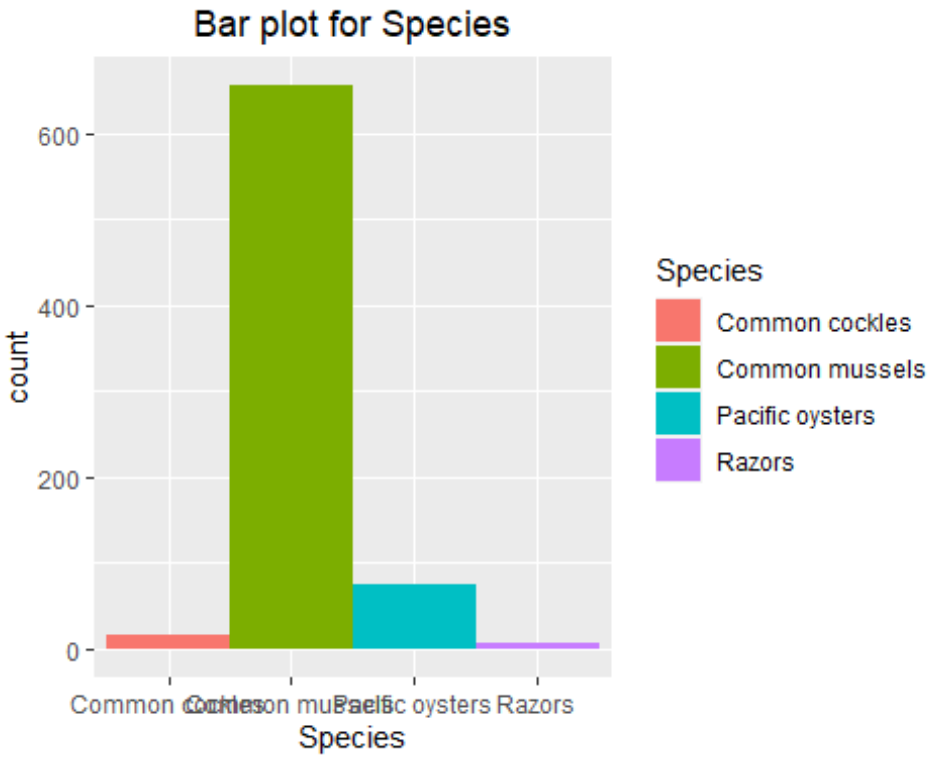


Positively skewed distribution because the long tail moves to the right. And the value is to the left of the mean. The mean is greater than the median. The mean is close to the right side of the distribution and the median is close to the left side of the distribution

QUESTION 5

Barplot to represent Data

```
p <- ggplot(plankton, aes(Species, fill=Species))
p <- p+ geom_bar(width = 1)
p <- p+ geom_bar(position = "stack")
p<- p+ labs(x="Species",
       title ="Bar plot for Species",)+
       theme(plot.title = element_text(hjust = 0.5))
p
```

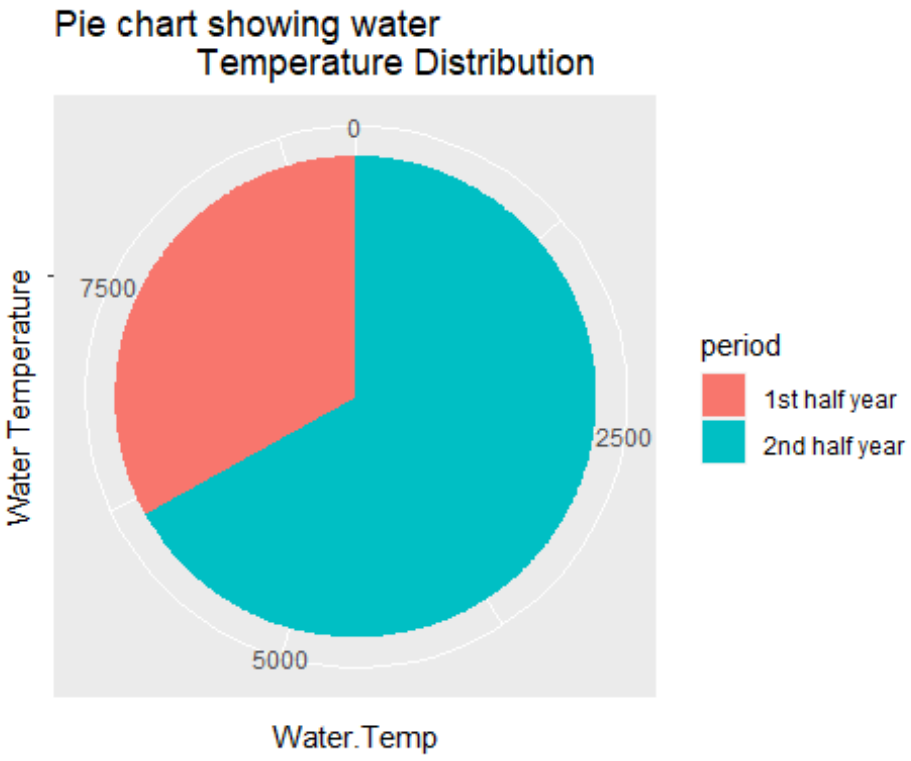


From the Bar graph, we can say that the Common mussels have the highest count in the species and the Razors with the lowest count.

```
QUESTION 6
```

plotting the pie chart for water temperature

```
p <- ggplot(plankton, aes(x="", y= Water.Temp, fill = period))
p <- p + geom_bar(width = 1, stat = "identity")
p <- p + coord_polar ("y", start = 0)
P <- p + theme_void()
p <- p + labs(x="Water Temperature", title = "Pie chart showing water
              Temperature Distribution")
p
```

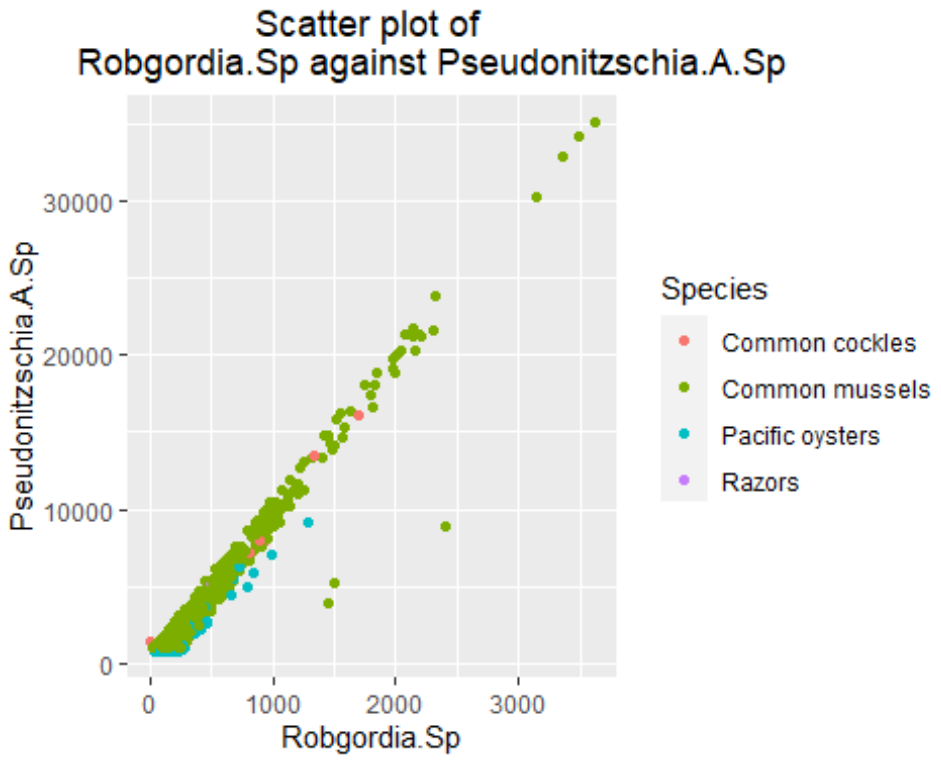


The pie chart shows that the water temperature is lower in the first half of the year and the highest in the second half of the year.

QUESTIION 7

##Plot showing values of Pseudonitzschia.A.Sp against values of Robgordia.S

```
p <- ggplot(plankton, aes(Robgordia.Sp,Pseudonitzschia.A.Sp, colour=Species))
p <- p + geom_point()
p<- p+ labs(x="Robgordia.Sp", y="Pseudonitzschia.A.Sp", title ="Scatter plot
of
            Robgordia.Sp against Pseudonitzschia.A.Sp",) +theme(plot.title =
element_text(hjust = 0.5))
p
```

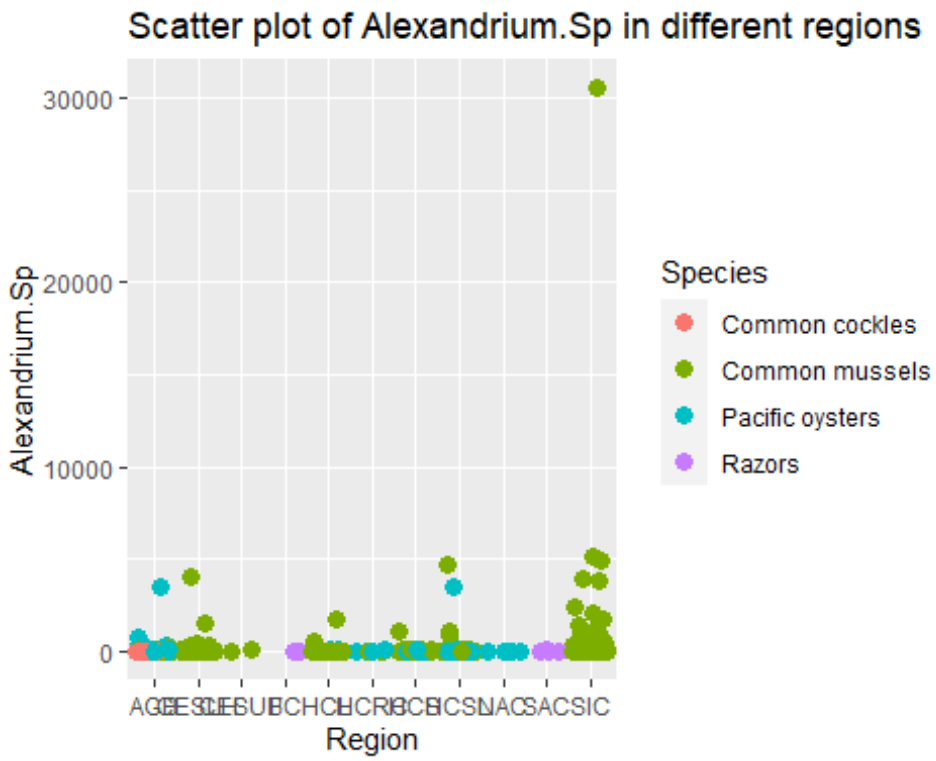


The values of Pseudonitzschia and Robgordia have a negative correlation because the value species Common mussels and pacific oysters are very close to the straight line. However, the three Common mussels points represent a negative residual, smaller than the predicted value.

QUESTION 8

Plotting a graph for of Alexandrium.Sp in different regions by species

```
p <- ggplot(plankton, aes(x = Region, y = Alexandrium.Sp, colour = Species))
p <- p + geom_point(position = "jitter", pch = 16, size = 3 )
p <- p + labs(y = "Alexandrium.Sp",
title ="Scatter plot of Alexandrium.Sp in different regions")
p
```



The plot above shows that there is a positive tendency among Alexandrium.Sp the species increase in different regions as Alexandrium.Sp value increases. Common mussels have an outlier furthest from the regression line, which is 3000. Adding jitter to the plot helps separate overlapping points to show the descriptive relationship.

Question 9

Discovering a pair of plankton species which are correlated and a pair which

```
cov(plankton$Pseudonitzschia.A.Sp,plankton$Robgordia.Sp)

## [1] 1931217

cov(plankton$Pseudonitzschia.A.Sp,plankton$Alexandrium.Sp)

## [1] 324833.9

cor(plankton$Pseudonitzschia.A.Sp,plankton$Robgordia.Sp)

## [1] 0.975273

cor(plankton$Pseudonitzschia.A.Sp,plankton$Alexandrium.Sp)

## [1] 0.06138349
```

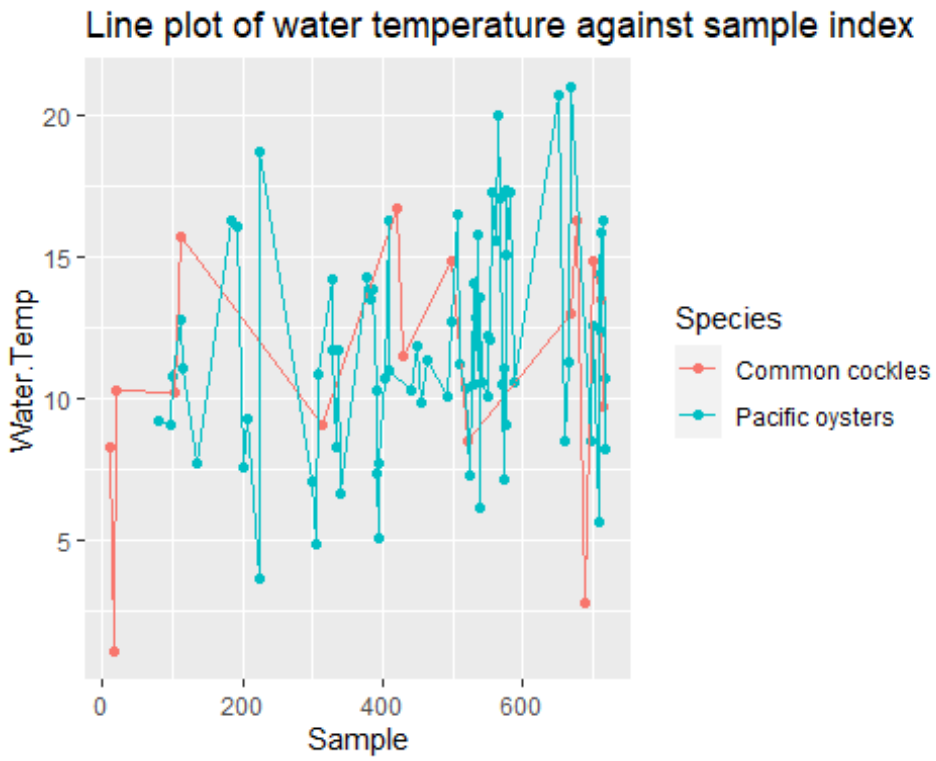
From the observation, it can be assume that pairs (Pseudonitzschia. A.Sp, Robgordia.Sp) has a strong correlation because their value is 0.975273 which is close to 1. While the other pair (Pseudonitzschia.A.Sp, Alexandrium.Sp) value is 0.06138349 which is not close to 1

and the covariance has a positive number hence there is no correlation between these pairs of species

QUESTION 10

Line plot of tempreture against Sample

```
plankton %>% filter(Species%in%c("Common cockles", "Pacific oysters")) %>%
ggplot(aes(x=Sample, y=Water.Temp, colour = Species)) +
geom_point() +
geom_line() +
labs(x = "Sample",
y = "Water.Temp",
title ="Line plot of water temperature against sample index")
```



QUESTION 11

##Producing a linear regression model of Pseudonitzschia.A.Sp on Robgordia.Sp ##forCommon mussels

```
 lm.output <- lm(formula = Pseudonitzschia.A.Sp ~ Robgordia.Sp, data =
plankton)
summary(lm.output)

##
## Call:
## lm(formula = Pseudonitzschia.A.Sp ~ Robgordia.Sp, data = plankton)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13696.2   -479.6     67.8    573.4   1999.9
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -275.97689   48.62444  -5.676 1.97e-08 ***
## Robgordia.Sp    9.50029    0.07851 121.014  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 971.3 on 752 degrees of freedom
## Multiple R-squared:  0.9512, Adjusted R-squared:  0.9511
## F-statistic: 1.464e+04 on 1 and 752 DF,  p-value: < 2.2e-16
```

The above statistics show that the R-squared value is 0.9511, which indicates the value of Pseudonitzschia.A.Sp and Robgordia are highly correlated, which shows a good model. The F-statistic value of 1.464e+04 is very significant for the p-value < 2.2e-16

Estimating the value of Pseudonitzschia.A.Sp for the value (1000,2500,4000)

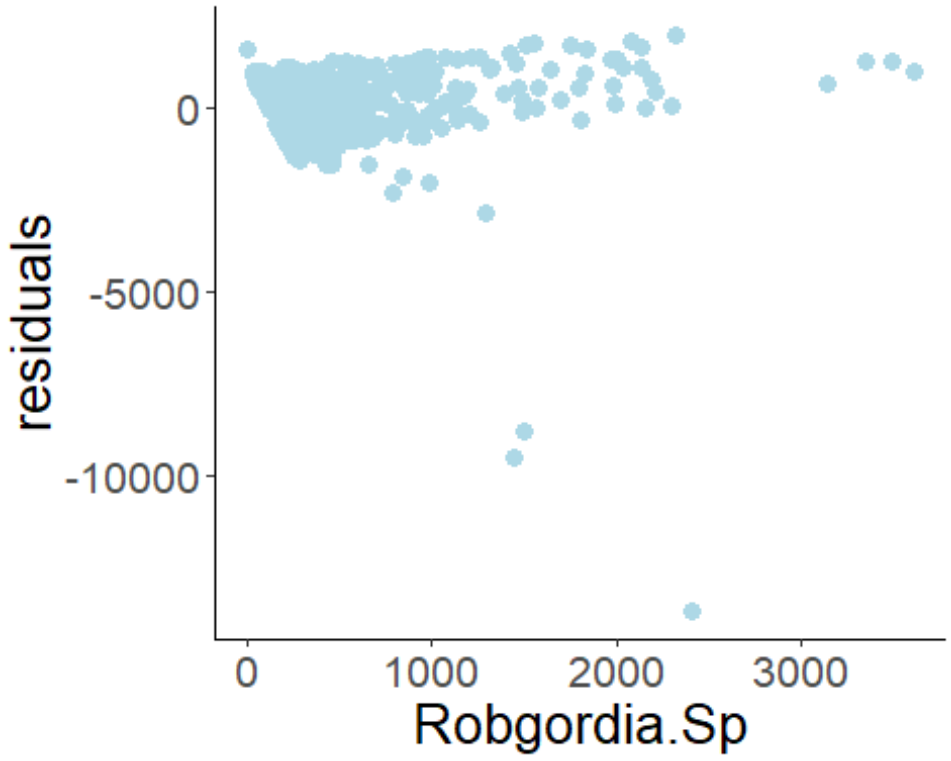
```
newdata <- data.frame(Robgordia.Sp=c(1000,2500,4000))
predict(lm.output,newdata)
```

```
##         1         2         3
##  9224.311 23474.744 37725.176
```

```
lmData <-
data.frame(residuals = lm.output$residuals,
                Pseudonitzschia.A.Sp=plankton$Pseudonitzschia.A.Sp,
                Robgordia.Sp= plankton$Robgordia.Sp)
```

ploting a graph for residuals

```
p <- ggplot(lmData, aes(x=Robgordia.Sp, y = residuals))
p <- p + geom_point(size=3, colour="lightblue")
p <- p +  theme_classic()
p <- p + theme(text = element_text(size = 20))
p
```

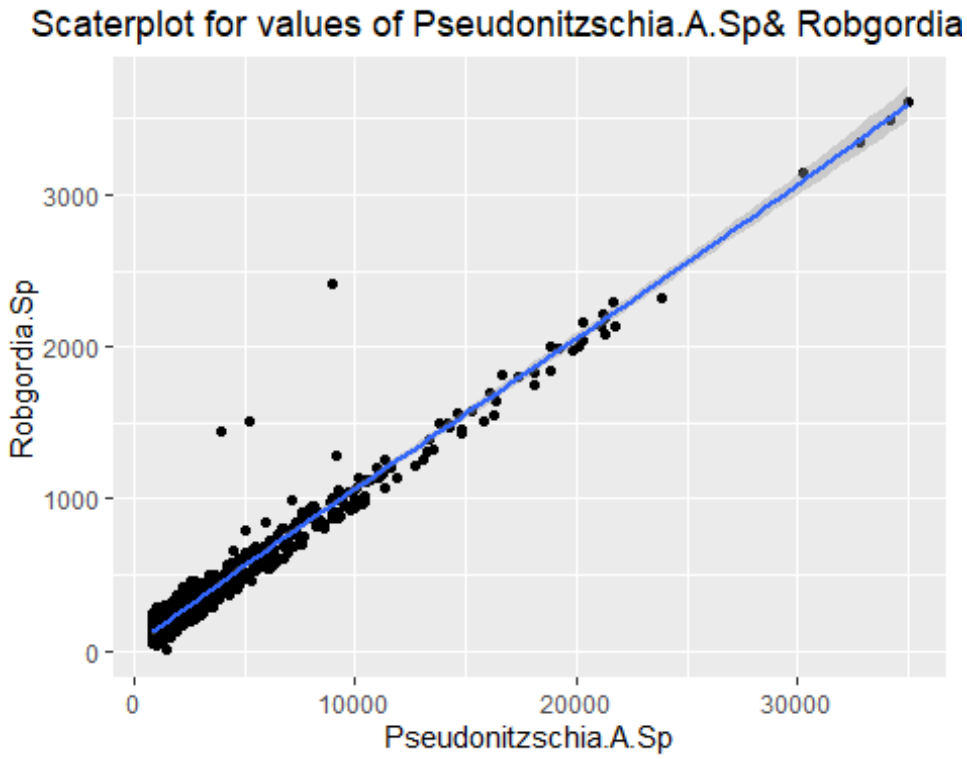


Using the spiro test to check the residual distribution is normal

```
shapiro.test(lmData$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  lmData$residuals
## W = 0.67413, p-value < 2.2e-16
```

Using the Shapiro-Wilk normality test, the p-value is smaller than 0.05. This value represents 0.000000000000000022, which is very close to zero. Therefore, the residuals are not considered to be normally distributed.

```
p <- ggplot(plankton,aes(x = Pseudonitzschia.A.Sp, y= Robgordia.Sp))
p <- p + geom_point() + stat_smooth()
p<- p+ labs(x="Pseudonitzschia.A.Sp",
       title ="Scaterplot for values of Pseudonitzschia.A.Sp& Robgordia.Sp
")+
       theme(plot.title = element_text(hjust = 0.5))

p

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



A linear regression model is significant because the residual scatterplots are randomly distributed and the graph above shows a linearly increasing relationship between Pseudonitzschia.A.Sp and Robgordia.Sp. However, the P-value for the Shapiro-Wilk normality test is less than 0.05, indicating that the null hypotheses is rejected, which is contradictory and that is my concern.

QUESTION 12

Observing if the temperature of the mean is 12 degrees at 99 confidence.

firstly create a dataframe with columns mean of the temperature of water,

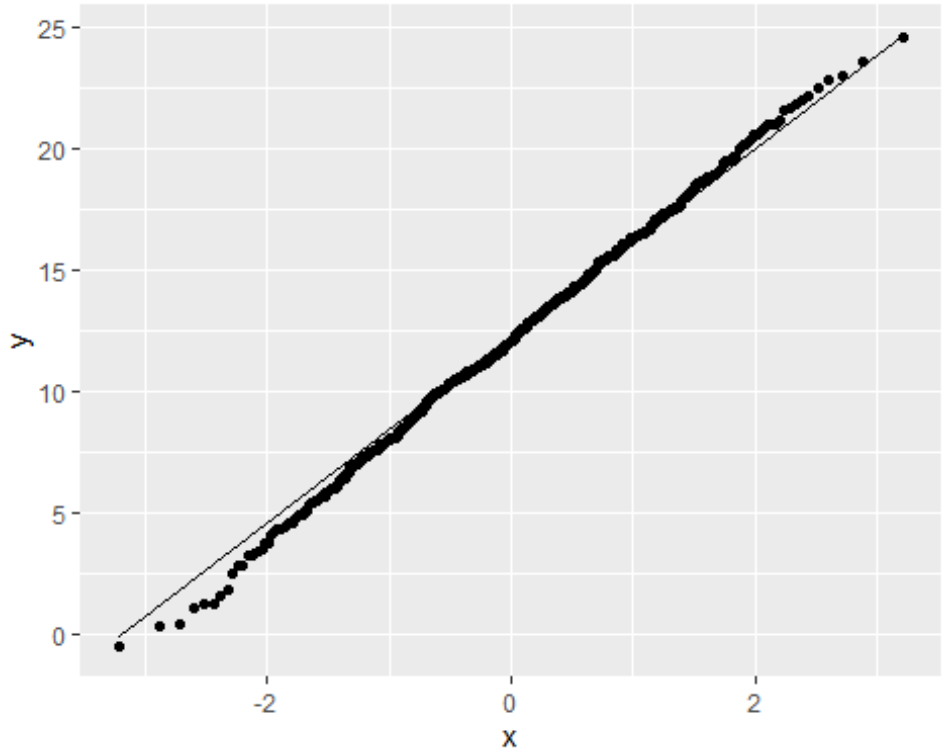
##month year period. ##creating the dataframe meanWater

```
newdata <- plankton[,c(5,10,11)]
meanWaterTemp<-aggregate(newdata$Water.Temp, by=list(newdata$month,
newdata$year),
FUN=mean, na.rm=TRUE)
colnames(meanWaterTemp) <- c("month","year","meanwatertemp")
View(meanWaterTemp)
summary(meanWaterTemp)

##      month             year      meanwatertemp
##  Min.   : 3.000   Min.   :2009   Min.   : 6.70
##  1st Qu.: 5.000   1st Qu.:2012   1st Qu.:10.81
##  Median : 7.000   Median :2015   Median :11.94
##  Mean   : 6.561   Mean   :2015   Mean   :12.13
```

```
##  3rd Qu.: 8.000   3rd Qu.:2018   3rd Qu.:13.60
##  Max.   :10.000   Max.   :2021   Max.   :17.10

CI(meanWaterTemp$meanwatertemp, ci=0.99)

##    upper     mean    lower
## 12.79912 12.13191 11.46471
```

At the 99% confidence interval, the mean water temperature is 12.13. So, we are 99% sure that the average temperature is 12 degrees. Without further testing, 95% were confident that the average temperature was not 12.5 degrees.

```
                              QUESTION 13

plankton$year.period <- plankton$period
plankton$year.period <- as.numeric(plankton$year.period)

p <- ggplot(plankton, aes(sample = Water.Temp))
p <- p + stat_qq()
p <- p + stat_qq_line()
p
```



Shapiro-walk normality test

```
shapiro.test(plankton$Water.Temp)

##
##  Shapiro-Wilk normality test
```

```
##
## data:  plankton$Water.Temp
## W = 0.9987, p-value = 0.869
```

It can be concluded that the dataset passes both tests. Firstly, in the QQ plot, the points are closer to a straight line, and the Shapiro-Wilks test shows a P-value greater than 0.05. This fact shows that the null hypothesis is true and the alternative hypothesis is false and there is a correlation between water temperature and the time of year. So, the suspicion is justified.

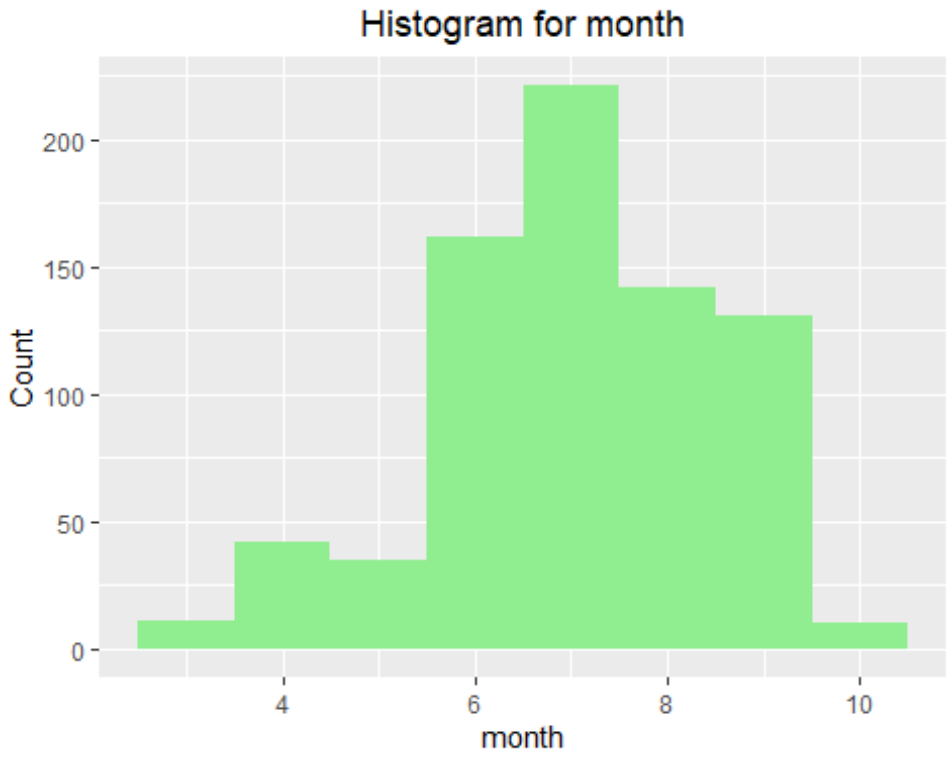
```
QUESTION 14
```

using ANOVA test to differentiate species formed in water temperature.

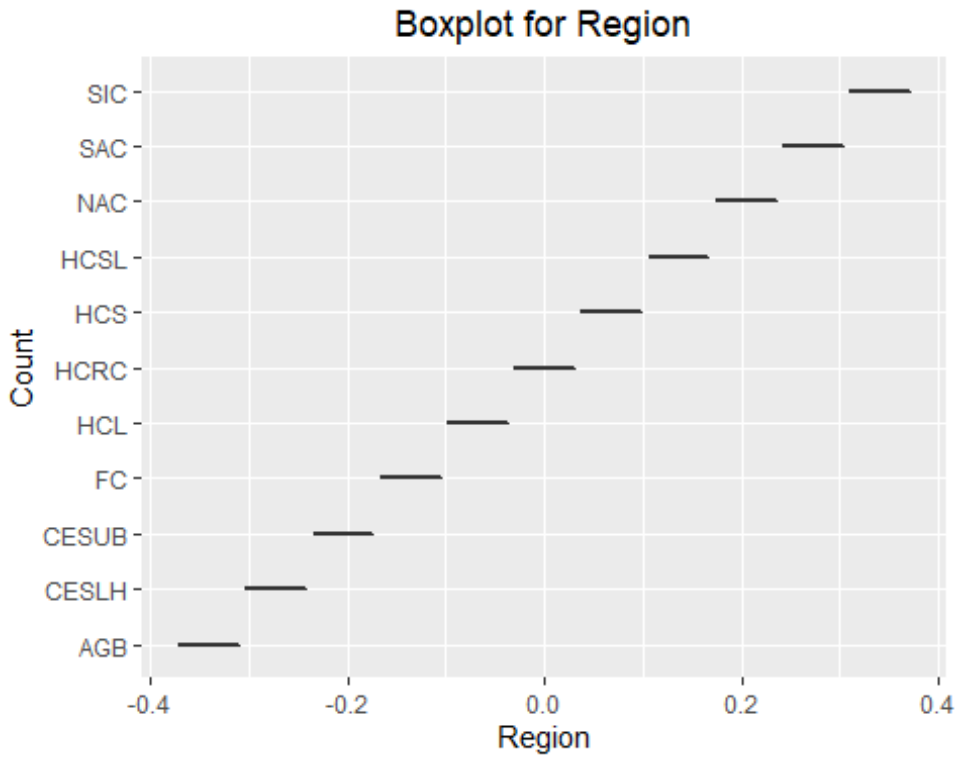
The ANOVA test tells us whether there is a difference between the means. So, to check for the validity of the Anova test, you apply the Q-Q plot test and the Shapiro-Wilk test. If the dot for the Q_Q plot is close to the line and the p-value for the Shapiro-Wilk test is higher than 0.05, It is reasonable to assume the distribution is normal. However, the ANOVA test's limitation is that it can only use to investigate a single variable. When distinguishing between the means of three or additional categories, it can tell us if at least one pair of means is significantly different, but not which pair it is.

```
QUESTION 15
```

```r
p <- ggplot(data =plankton, aes(month))
p <- p + geom_histogram( fill="lightgreen",binwidth=1)
p <- p + labs(x="month",
              y="Count",
              title = "Histogram for month")+
              theme(plot.title = element_text(hjust = 0.5))
p
```




```
p <- ggplot(plankton, aes(y=Region))
p <- p + geom_boxplot(fill="lightgreen",)
p <- p + labs(x="Region",
         y="Count",
         title = "Boxplot for Region")+
         theme(plot.title = element_text(hjust = 0.5))
p
```



From the above graph, we can observe that AGB is the area with the least counts and SIC is the area with the most counts. While for the other plot, the half of the month has the highest count and decreases the count at the start and end of the month. Comparing this plot, we can assume that the histogram is negatively skewed and that the bar graph increase along each region. The count increases as data move from one region to another. Although there is an undercount in the data from one month to another