ASSIGMENT

ENGR. SAMUEL EFFIONG

3/20/2022

# Solution To Assigntment

Analysis and visual representation of **plankton** in water where shellfish are farmed DATASET was gotten from <plankton.csv>

## Plankton Dataset set for the project.

# SOLUTION 1

* Using **Univariate statistics** to analyse the **plankton dataset**
* FOR **MEAN,MEDIAN and MODE** in summary statistics

Table continues below

| PLANKTONDATASET | Pseudonitzschia.A.Sp | Alexandrium.Sp | Robgordia.Sp |
| --- | --- | --- | --- |
| RESULT | 3762 | 146 | 425 |

| Water.Temp |
| --- |
| 12.17 |

Table continues below

| PLANKTONDATASET | Pseudonitzschia.A.Sp | Alexandrium.Sp | Robgordia.Sp |
| --- | --- | --- | --- |
| RESULT | 2180 | 0 | 276.6 |

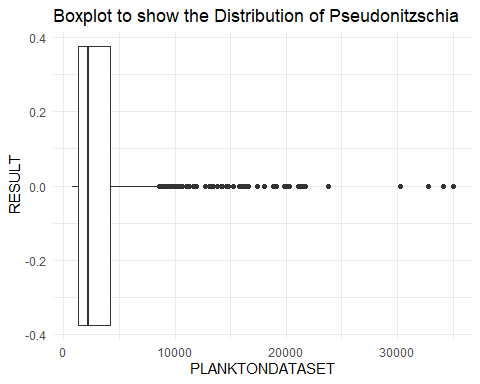
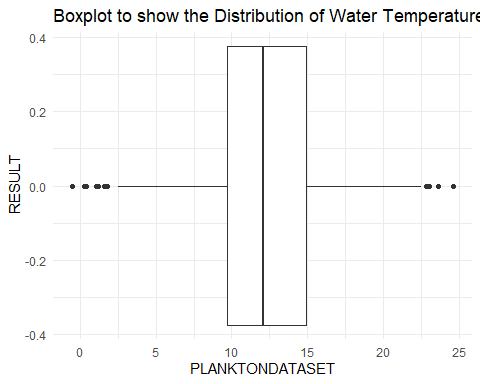
| Water.Temp |
| --- |
| 12.1 |

Table continues below

| PLANKTONDATASET | Pseudonitzschia.A.Sp | Alexandrium.Sp | Robgordia.Sp |
| --- | --- | --- | --- |
| RESULT | 4392 | 1205 | 450.9 |

| Water.Temp |
| --- |
| 4.123 |

# SOLUTION 2

* A **Boxplot** to show the distribution of **Pseudonitzschia** and a second one to show the distribution of **water temperature.**
* **FOR PSEUDONITZSCHIA** 
* **FOR WATER TEMPERATURE** 

### COMMENT

* In the **Pseudonitzschia.A.Sp** it was observe that there were more, *extreme outliers* than in the **water Temperature.**
* IN the **water temperature** plot it was observe that it was a *normal plot.*

# SOLUTION 3

* Using **Univariate statistics** to compare data for **Pseudonitzschia in the year**, 2021 with its data in previous years.

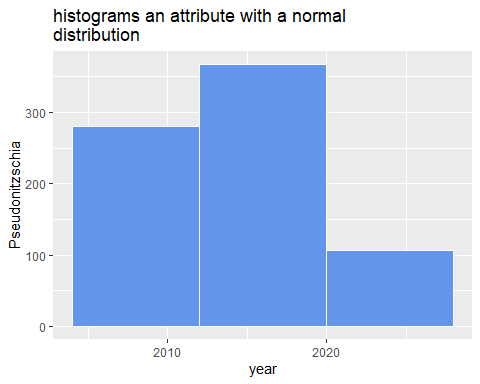
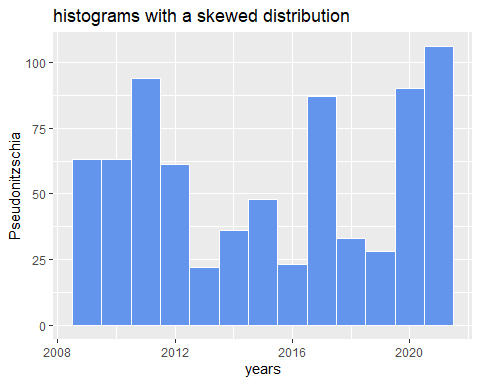
| year | Pseudonitzschia.A.Sp |
| --- | --- |
| 2021 | 106 |
| 2011 | 94 |
| 2020 | 90 |
| 2017 | 87 |
| 2009 | 63 |
| 2010 | 63 |
| 2012 | 61 |
| 2015 | 48 |
| 2014 | 36 |
| 2018 | 33 |
| 2019 | 28 |
| 2016 | 23 |
| 2013 | 22 |

### COMMENT

* form the **univarite data** above,There was great improvement in the year 2021 compared to other years.

# SOLUTION 4

* Histogram difference between **skewed distribution and normal distribution**



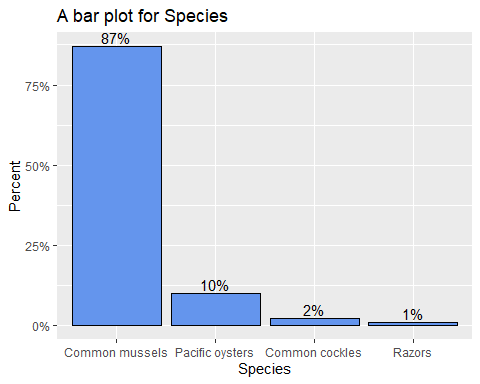
### COMMENT

\*Skewed distribution was not evenly distributed.hence we can’t get a symmetrical curve.

\*Normal distribution was evenly distributed.hence we get a symmetrical curve.

# SOLUTION 5

* A bar plot for **Species**

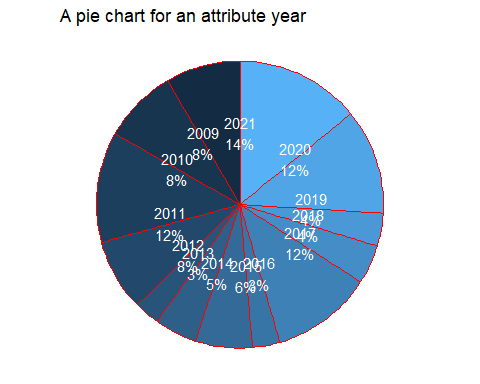


### COMMENT

* I choose to use Species because it has different category.
* There are are wide difference between all of the category and **Common mussels** was observe to be the highest.

# SOLUTION 6

* A **pie chart** for an attribute of your choice

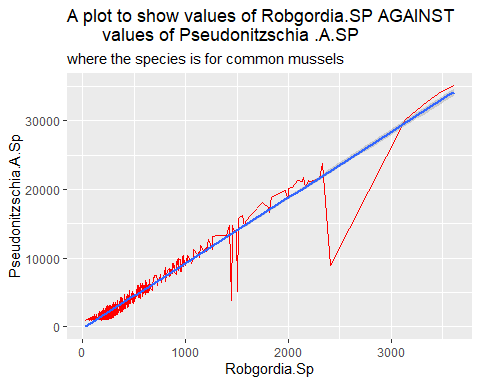


### COMMENT

* I observe in the graph that that some years have smaller percentage than others. thus there is difference in **planktondata** of each year.

# SOLUTION 7

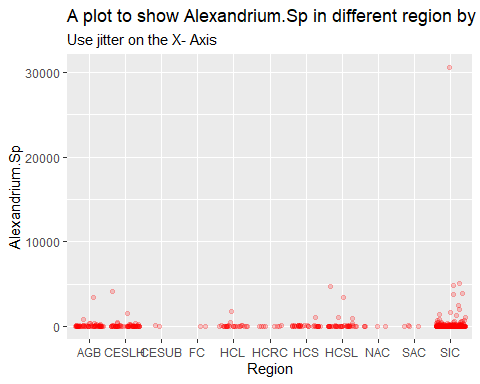
* A plot to show values of **Robgordia.SP** AGAINST values of ,**Pseudonitzschia .A.SP** where the species is either common mussels, or pacific oysters.



### COMMENT

# SOLUTION 8

* A plot to show **Alexandrium.Sp** in different region by **farming species.**



### COMMENT

* I observe that common common mussels values decrease and increase at some point. this go a long way to tell us that there are great significant difference between the two planktondata.

# SOLUTION 9

* To determine the pair that our **Correlated and non Correlated** we first:
  + Testing the assumptions (Linearity and Normalcy) with SHAPIRo - WILK
  + Correlation Coefficient
  + Checking for the significance

##   
## Shapiro-Wilk normality test  
##   
## data: planktondata$Pseudonitzschia.A.Sp  
## W = 0.61805, p-value < 2.2e-16

##   
## Shapiro-Wilk normality test  
##   
## data: planktondata$Alexandrium.Sp  
## W = 0.083029, p-value < 2.2e-16

##   
## Shapiro-Wilk normality test  
##   
## data: planktondata$Robgordia.Sp  
## W = 0.65745, p-value < 2.2e-16

##   
## Shapiro-Wilk normality test  
##   
## data: planktondata$Pseudonitzschia.A.Sp  
## W = 0.61805, p-value < 2.2e-16

## [1] 0.06138349

## [1] 0.975273

## [1] 0.06138349

## [1] 0.06482684

## [1] 0.975273

## [1] 0.06482684

##   
## Pearson's product-moment correlation  
##   
## data: planktondata$Pseudonitzschia.A.Sp and planktondata$Alexandrium.Sp  
## t = 1.6865, df = 752, p-value = 0.09212  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.01005901 0.13220249  
## sample estimates:  
## cor   
## 0.06138349

##   
## Pearson's product-moment correlation  
##   
## data: planktondata$Pseudonitzschia.A.Sp and planktondata$Robgordia.Sp  
## t = 121.01, df = 752, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.9715248 0.9785333  
## sample estimates:  
## cor   
## 0.975273

##   
## Pearson's product-moment correlation  
##   
## data: planktondata$Alexandrium.Sp and planktondata$Pseudonitzschia.A.Sp  
## t = 1.6865, df = 752, p-value = 0.09212  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.01005901 0.13220249  
## sample estimates:  
## cor   
## 0.06138349

##   
## Pearson's product-moment correlation  
##   
## data: planktondata$Alexandrium.Sp and planktondata$Robgordia.Sp  
## t = 1.7815, df = 752, p-value = 0.07524  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.006602121 0.135597630  
## sample estimates:  
## cor   
## 0.06482684

##   
## Pearson's product-moment correlation  
##   
## data: planktondata$Robgordia.Sp and planktondata$Pseudonitzschia.A.Sp  
## t = 121.01, df = 752, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.9715248 0.9785333  
## sample estimates:  
## cor   
## 0.975273

##   
## Pearson's product-moment correlation  
##   
## data: planktondata$Robgordia.Sp and planktondata$Alexandrium.Sp  
## t = 1.7815, df = 752, p-value = 0.07524  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.006602121 0.135597630  
## sample estimates:  
## cor   
## 0.06482684

## OR WE COULD DO ALL THREE TOGETHER

## Pseudonitzschia.A.Sp Robgordia.Sp Alexandrium.Sp  
## Pseudonitzschia.A.Sp 1.00000000 0.97527304 0.06138349  
## Robgordia.Sp 0.97527304 1.00000000 0.06482684  
## Alexandrium.Sp 0.06138349 0.06482684 1.00000000

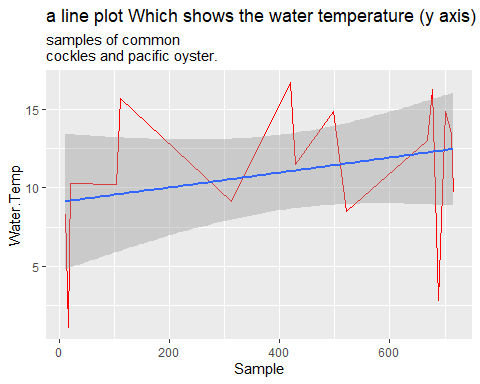
## Pseudonitzschia.A.Sp Robgordia.Sp Alexandrium.Sp   
## Min. :0.06138 Min. :0.06483 Min. :0.06138   
## 1st Qu.:0.51833 1st Qu.:0.52005 1st Qu.:0.06311   
## Median :0.97527 Median :0.97527 Median :0.06483   
## Mean :0.67889 Mean :0.68003 Mean :0.37540   
## 3rd Qu.:0.98764 3rd Qu.:0.98764 3rd Qu.:0.53241   
## Max. :1.00000 Max. :1.00000 Max. :1.00000

### CONCLUSION and Justify

* Since the p-value is above 0.05 (here it is 0.09212, 2.2e-16,0.09212, and 0.07524, we can conclude that **Pseudonitzschia.A.Sp**, **Alexandrium.Sp** and **Robgordia.Sp** are significantly correlated with a value of 0.06482684,0.975273 and 0.06138349 and a p-value of 0.09212,2.2e-16,0.09212, and 0.07524.
* As we can see there is a positive value between **Pseudonitzschia.A.Sp**, **Alexandrium.Sp** and **Robgordia.Sp**, the point to be noted here is correlation is just a measure of association.
* It will tell the degree of association along with the direct or indirect proportionality.
* since all three pairing are above 0.05 , we thus conclude that the is no pair among planktondata( Pseudonitzschia.A.Sp Alexandrium.Sp Robgordia.Sp)

# SOLUTION 10

* A line plot Which shows the **water temperature** (y axis) against the **sample index** (x axis), for *samples of common cockles and pacific oyster*.



#SOLUTION 11

* linear regression model of **Pseudonsitzshia.A.SP** for a value of **Robgordia.sp** of 1000,2500,and 4000 *cells per liter*.

## Pseudonitzschia.A.Sp Robgordia.Sp  
## Pseudonitzschia.A.Sp 1.000000 0.975273  
## Robgordia.Sp 0.975273 1.000000

##   
## Call:  
## lm(formula = Pseudonitzschia.A.Sp ~ Robgordia.Sp, data = data101)  
##   
## 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

## [1] 9224.313

## [1] 23474.75

## [1] 37725.18

### justify the appropriateness of the model and comments

* We found a significant relationship between **Pseudonsitzshia.A.SP** and **Robgordia.sp** (p < 2.2e-16, R2 = 0.9512), with a **Robgordia.sp** of 1000,2500,and 4000 cells per liter there is *increase*.

# SOLUTION 12

* A data frame with three columns: month,year,and the mean of the **water temperatures observed in the planktondataset** during that month-year period

## Average.temp Month Year  
## 1 13.766667 10 2010  
## 2 14.950000 10 2013  
## 3 16.200000 10 2014  
## 4 11.000000 10 2015  
## 5 13.650000 10 2016  
## 6 9.900000 3 2011  
## 7 14.800000 3 2013  
## 8 6.700000 3 2015  
## 9 10.800000 3 2016  
## 10 11.100000 3 2019  
## 11 13.533333 3 2020  
## 12 10.800000 3 2021  
## 13 14.500000 4 2009  
## 14 10.371429 4 2011  
## 15 7.133333 4 2012  
## 16 9.960000 4 2014  
## 17 12.533333 4 2017  
## 18 14.900000 4 2018  
## 19 13.550000 4 2019  
## 20 13.855556 4 2020  
## 21 15.383333 4 2021  
## 22 16.600000 5 2009  
## 23 17.100000 5 2010  
## 24 15.855556 5 2012  
## 25 12.300000 5 2014  
## 26 7.400000 5 2016  
## 27 15.433333 5 2017  
## 28 15.500000 5 2018  
## 29 10.133333 5 2019  
## 30 10.500000 5 2020  
## 31 12.650000 5 2021  
## 32 11.811765 6 2009  
## 33 12.221053 6 2010  
## 34 10.946154 6 2011  
## 35 10.894737 6 2012  
## 36 16.600000 6 2013  
## 37 9.736364 6 2014  
## 38 11.677778 6 2015  
## 39 9.525000 6 2016  
## 40 11.975000 6 2017  
## 41 11.705882 6 2018  
## 42 8.140000 6 2019  
## 43 11.833333 6 2020  
## 44 14.112500 6 2021  
## 45 12.856522 7 2009  
## 46 13.518182 7 2010  
## 47 11.560870 7 2011  
## 48 12.371429 7 2012  
## 49 9.144444 7 2013  
## 50 13.280000 7 2014  
## 51 13.000000 7 2015  
## 52 7.466667 7 2016  
## 53 12.891111 7 2017  
## 54 10.825000 7 2018  
## 55 12.822222 7 2019  
## 56 12.676000 7 2020  
## 57 11.086207 7 2021  
## 58 11.750000 8 2009  
## 59 11.411111 8 2010  
## 60 12.680488 8 2011  
## 61 13.620000 8 2012  
## 62 10.285714 8 2013  
## 63 10.775000 8 2014  
## 64 11.406250 8 2015  
## 65 13.150000 8 2016  
## 66 10.725000 8 2017  
## 67 8.000000 8 2018  
## 68 15.775000 8 2019  
## 69 11.035294 8 2020  
## 70 12.175000 8 2021  
## 71 11.566667 9 2009  
## 72 13.360000 9 2010  
## 73 11.411111 9 2011  
## 74 14.675000 9 2012  
## 75 8.450000 9 2013  
## 76 11.914286 9 2014  
## 77 11.460000 9 2015  
## 78 13.775000 9 2016  
## 79 12.360000 9 2017  
## 80 9.700000 9 2018  
## 81 13.955000 9 2020  
## 82 11.863636 9 2021

### Water temperature is 12 deg

* H0 -> mu = 12 (the water temperature is 12 degree)
* H1 -> mu != 12 (the water temperature is not 12 degree)

##   
## One Sample t-test  
##   
## data: Water.Temp  
## t = 1.1563, df = 753, p-value = 0.2479  
## alternative hypothesis: true mean is not equal to 12  
## 99 percent confidence interval:  
## 11.78588 12.56134  
## sample estimates:  
## mean of x   
## 12.17361

* From the obtained P-value there is no enough statistical evidence to reject H0 which state that the **Water temperature is 12 degrees**
* Since at 99% level of confidence the estimated mean of the **water temperature** is between 11.8 and 12.6 12.5, should be at the *extreme of the sampling distribution of the mean*
* So if 12.5 should be moved to the middle of the distribution which will distort the confidence interval making the mean value 12.17 far off to the left of the **sampling distribution**
* Therefore when a using type I error criteria of 95% we’ll obtain a low p-value making it easier to reject the null *hypothesis* which will come to a conclusion of the **water temperature** not being 12.5

# SOLUTION 13

* water temperature affected period
* H0 -> mu1 == mu2 (Water temperature of 1st half period is equal to **water temperature** of the 2nd half period)
* H1 -> mu1 != mu2 (Water temperature of 1st half period is not equal to **water temperature** of the 2nd half period)

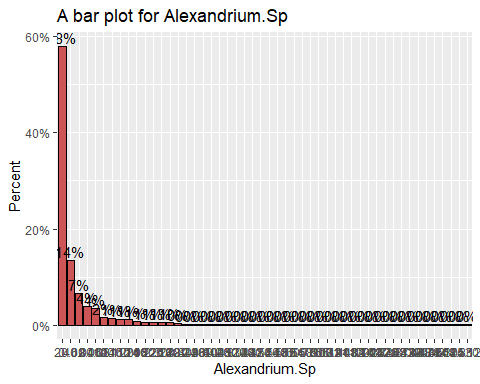
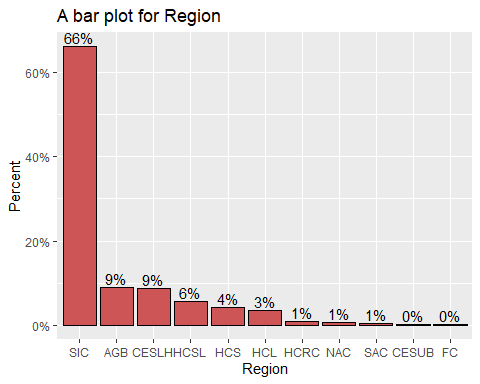
##   
## Welch Two Sample t-test  
##   
## data: Water.Temp by period  
## t = -0.22595, df = 451.47, p-value = 0.8213  
## alternative hypothesis: true difference in means between group 1st half year and group 2nd half year is not equal to 0  
## 99 percent confidence interval:  
## -0.9312907 0.7816653  
## sample estimates:  
## mean in group 1st half year mean in group 2nd half year   
## 12.12360 12.19841

* From the above analysis its clear that there is no difference between the **water temperature** in both the *1st and 2nd half of the year*, so we conclude that the **water temperature** is not affected by the period

# SOLUTION 14

* Our work is not done once we have found a statistically significant difference between the **group means**. When we calculate , and end up with an average difference between the group means.
* Since we are comparing four group means, we might find a relatively large average difference between these group means even if two of the four **group means are identical**.
* Therefore,a statistically significant value tells us only that somewhere there is a meaningful difference between the *group means*. But it does not tell us which **groups differ** from each other significantly.

# SOLUTION 15

* FOR EFFECTIVE PLOT AND NON EFFECTIVE 

### COMMENT

* the first graph can easily be read. while the second can be read but parameter and measurement are not clear.

# Rferences

* Univariate Table: <https://www.rdocumentation.org/packages/Publish/versions/2020.12.23/topics/univariateTable>
* DATA VISUALIZATION by R studio: <https://rstudio-pubs-static.s3.amazonaws.com/294957_a9f61a4e682743daa04390484e6e5a7b.html#bar_plots>
* DATA VISUALIZATION WITH R 2: <https://rkabacoff.github.io/datavis/Univariate.html>
* Univariate and bivariate descriptive analysis BY:Gert Janssenswillen <https://beta.rstudioconnect.com/content/3350/dplyr_tutorial.html>
* [corollated https://cran.r-project.org/web/packages/psych/vignettes/intro.pdf](corollated%20https://cran.r-project.org/web/packages/psych/vignettes/intro.pdf)
* Stack Overflow:<https://stackoverflow.com/questions/64992231/two-ggplot-with-subset-in-pipe>