Proyecto personal de análisis de datos

Juan José Rodríguez Maulén

30 June 2022

Evaluación de la Madurez en Salmon Coho (Coho Salmon (COS))

Descripción del problema a analizar Se observó madurez en peces cultivados en la región de Los Lagos, sometidos a regimen de fotoperiodo para prevenir la madurez. El proveedor de fotoperiodo correspondió a la empresa BIOLED quienes utilizaron 3 intensidades lumínicas (W) en centros de cultivos de peces provenientes de las pisciculturas Huincara, Coipue, Lican y Lago Rupanco. En terminos de madurez observada por mix de jaulas, estas fluctuaron entre un 0 a un 20%, mientras que a nivel de centro de cultivo, esta alcanzo un 6,81% siendo aceptable un 5%.

Descripción de las variables de estudio, factores a analizar y el número total de observaciones Los datos de madurez, correspondieron a las observaciones realizadas en plantas de proceso, para la clasificación de calidades, donde una de las causales de degradación correspondio a madurez por jaula y centro de cultivo

Variable respuesta= % de maduración (Cuantitativa Discreta) Variable explicativa= Centro de Cultivo (Cualitativa Nominal) y Potencia (W) (Cuantitativa Discreta) N= 220 observaciones.

Hipotesis de investigación: Centros con fotoperiodo no presentan diferencias significativas en el desarrollo de madurez respecto a centros sin fotoperiodo

Utiliza paquetes para importar y analizar datos.

```
knitr::opts_chunk$set(echo = TRUE)
library(datasets)
library(ggplot2)
library(readxl)
library(stats)
library(lme4)
```

```
## Loading required package: Matrix
```

```
library(Matrix)
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(readr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
```

```
## v purrr   0.3.4     v forcats 0.5.1
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x psych::%+%()    masks ggplot2::%+%()
## x psych::alpha()  masks ggplot2::alpha()
## x tidyr::expand() masks Matrix::expand()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x tidyr::pack()   masks Matrix::pack()
## x tidyr::unpack() masks Matrix::unpack()
```

```
library(janitor)
```

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

```
library(scales)
```

```
##
## Attaching package: 'scales'
## The following object is masked from 'package:purrr':
##
##     discard
## The following object is masked from 'package:readr':
##
##     col_factor
## The following objects are masked from 'package:psych':
##
##     alpha, rescale
```

```
library(ggthemes)
library(ggrepel)
library(xlsx)
library(gridExtra)
```

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

```
Madurez <- read_excel("/cloud/project/Coho Season 2021-2022.xlsx",na="NA", sheet = 2)
Madurez <- na.omit(Madurez)

#############################################################################################
#Madurez$JAULA <- as.factor (Madurez$JAULA)
#Madurez$Centro2 <- as.factor (Madurez$Centro2)
#Madurez$SEXO <- as.factor (Madurez$SEXO)
#Madurez$`TIPO GRUPO` <- as.factor (Madurez$`TIPO GRUPO`)
#Madurez$Tipo_Sexo <- as.factor (Madurez$Tipo_Sexo)
#Madurez$`Jaula individual`<-as.factor(Madurez$`Jaula individual`)
```

```
#Madurez$Fotoperiodo <- as.factor(Madurez$Fotoperiodo)
```

Resumen

Para obtención de estadística descriptiva

```
summary(Madurez)
```

```
##  Jaula individual  Site Name            Cage           % Deformation
##  Min.   :0.0000   Length:79          Length:79          Min.   :0.0003672
##  1st Qu.:1.0000   Class :character   Class :character   1st Qu.:0.0174165
##  Median :1.0000   Mode  :character   Mode  :character   Median :0.0285547
##  Mean   :0.9873                                         Mean   :0.0325549
##  3rd Qu.:1.0000                                         3rd Qu.:0.0446231
##  Max.   :1.0000                                         Max.   :0.1071461
##     %Mature         % Desadaptado        Origen          Tipo orgien
##  Min.   :0.00000   Min.   :0.00000   Length:79          Length:79
##  1st Qu.:0.00000   1st Qu.:0.00355   Class :character   Class :character
##  Median :0.00998   Median :0.01276   Mode  :character   Mode  :character
##  Mean   :0.02628   Mean   :0.02100
##  3rd Qu.:0.04690   3rd Qu.:0.03070
##  Max.   :0.16881   Max.   :0.11803
##   Fotoperiodo      Proveedor            Potencia
##  Min.   :0.0000   Length:79          Min.   :   0
##  1st Qu.:1.0000   Class :character   1st Qu.:1800
##  Median :1.0000   Mode  :character   Median :2400
##  Mean   :0.7595                      Mean   :1808
##  3rd Qu.:1.0000                      3rd Qu.:2400
##  Max.   :1.0000                      Max.   :3600
```

```
head(Madurez)
```

```
## # A tibble: 6 x 11
##   `Jaula individual` `Site Name` Cage  `% Deformation` `%Mature` `% Desadaptado`
##                <dbl> <chr>       <chr>           <dbl>     <dbl>           <dbl>
## 1                  1 Teupa       102            0.0680    0              0.0109
## 2                  1 Teupa       103            0.0451    0              0.00675
## 3                  1 Teupa       104            0.0464    0              0.00766
## 4                  1 Teupa       107            0.0692    0              0.0102
## 5                  1 Punta Yoye  116            0.0317    0.0149         0.0743
## 6                  1 Punta Yoye  201            0.0452    0              0.0416
## # ... with 5 more variables: Origen <chr>, `Tipo orgien` <chr>,
## #   Fotoperiodo <dbl>, Proveedor <chr>, Potencia <dbl>
```

```
names(Madurez)
```

```
##  [1] "Jaula individual" "Site Name"        "Cage"             "% Deformation"
##  [5] "%Mature"          "% Desadaptado"    "Origen"           "Tipo orgien"
##  [9] "Fotoperiodo"      "Proveedor"        "Potencia"
```

Resumen de los datos

Categorias definidad por: Jaula, Site name, Cage, Origen, Proveedor, Potencia

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

```
## tibble [79 x 11] (S3: tbl_df/tbl/data.frame)
##  $ Jaula individual: num [1:79] 1 1 1 1 1 1 1 1 1 ...
##  $ Site Name       : chr [1:79] "Teupa" "Teupa" "Teupa" "Teupa" ...
##  $ Cage            : chr [1:79] "102" "103" "104" "107" ...
##  $ % Deformation   : num [1:79] 0.068 0.0451 0.0464 0.0692 0.0317 ...
##  $ %Mature         : num [1:79] 0 0 0 0 0.0149 ...
##  $ % Desadaptado   : num [1:79] 0.01085 0.00675 0.00766 0.01021 0.07434 ...
##  $ Origen          : chr [1:79] "HUINCACARA" "HUINCACARA" "HUINCACARA" "HUINCACARA" ...
##  $ Tipo orgien     : chr [1:79] "Piscicultura" "Piscicultura" "Piscicultura" "Piscicultura" ...
##  $ Fotoperiodo     : num [1:79] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Proveedor       : chr [1:79] "No" "No" "No" "No" ...
##  $ Potencia        : num [1:79] 0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, "na.action")= 'omit' Named int [1:141] 5 6 7 8 9 10 13 14 15 16 ...
##   ..- attr(*, "names")= chr [1:141] "5" "6" "7" "8" ...
```

Evaluación balanceo de datos Site Name

Para el caso de la categoría Site name, se puede observar que los niveles “Chidhuapi 1”; “Chope” y “Colaco 4” tienen el mayor número de observaciones, mientras que los demas niveles, presentan menos de cinco observaciones, por lo tanto, no está totalmente balanceado.

```
knitr::opts_chunk$set(echo = TRUE)
table(Madurez$`Site Name`)
```

```
##
## Chidhuapi 1 Chidhuapi 2 Chidhuapi 3       Chope    Colaco 4       Compu
##          18           8           4          16          18           2
##     Linguar   Malomacum  Punta Yoye       Teupa
##           2           5           2           4
```

Evaluación balanceo de datos Proveedor

Para el caso de la categoría Proveedor, se puede observar que el nivel “Bioled” tiene mayor número de observaciones, por lo tanto, no está totalmente balanceado.

```
knitr::opts_chunk$set(echo = TRUE)
table(Madurez$Proveedor)
```

```
##
## Bioled     No
##     60     19
```

Evaluación balanceo de datos Tipo Origen

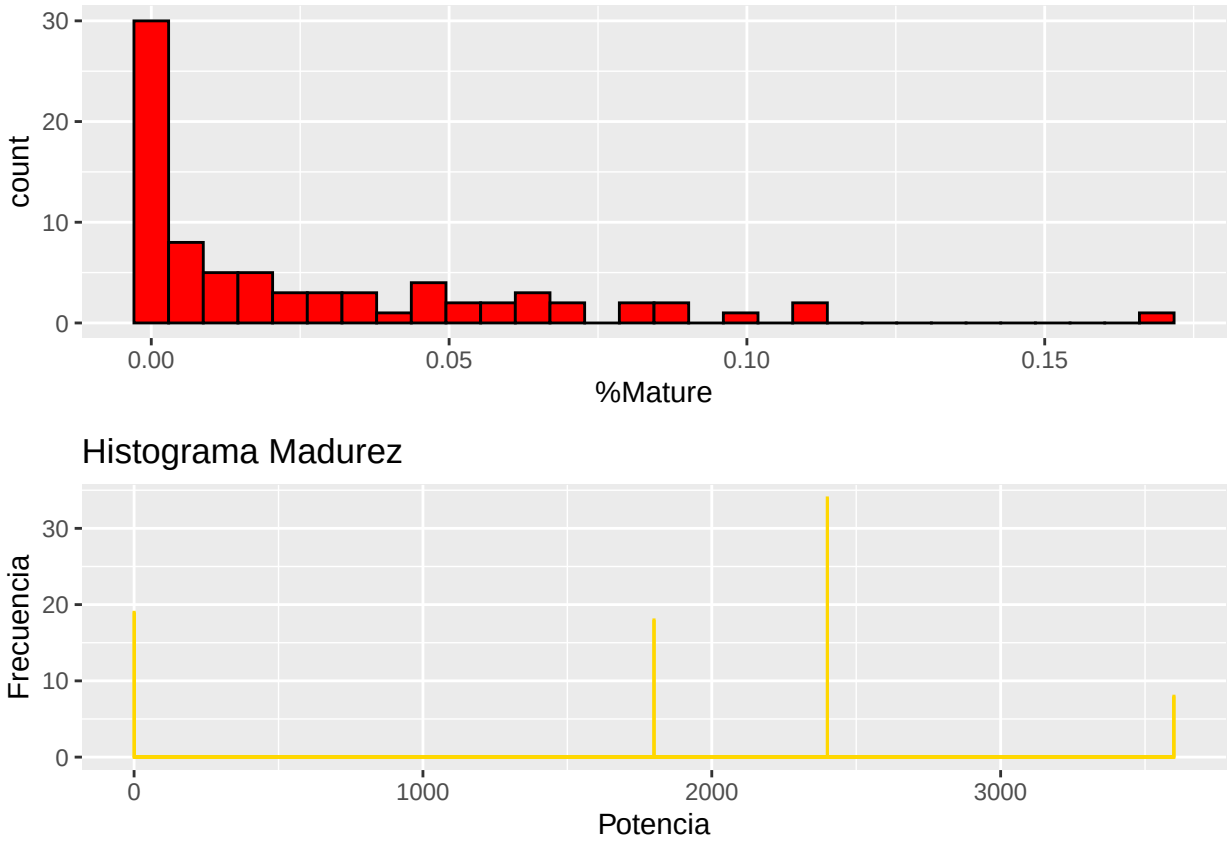
Para el caso de la categoría Tipo Origen, se puede observar que el nivel “Lago” tiene mayor número de observaciones respecto a Pscicultura, por lo tanto, no está totalmente balanceado.

```
knitr::opts_chunk$set(echo = TRUE)
table(Madurez$`Tipo orgien`)
```

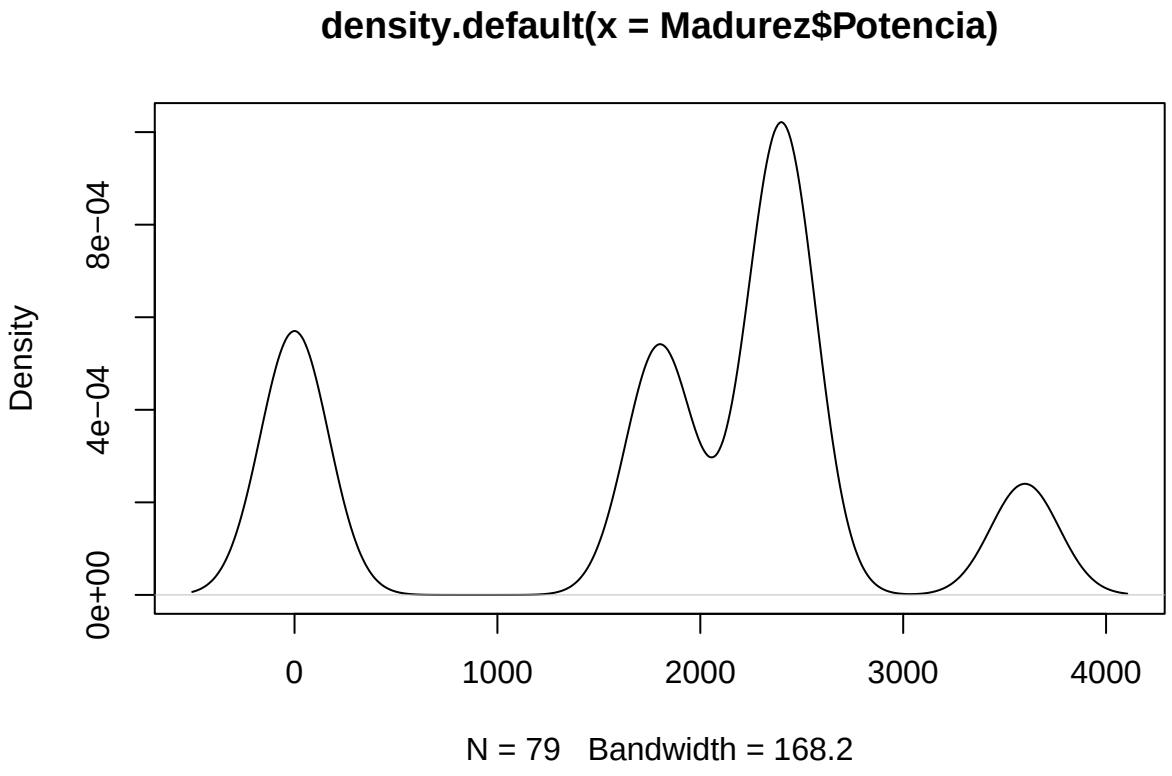
```
##
##         Lago Piscicultura
##           62           17
```

Describa la variación de las variables usando histogramas

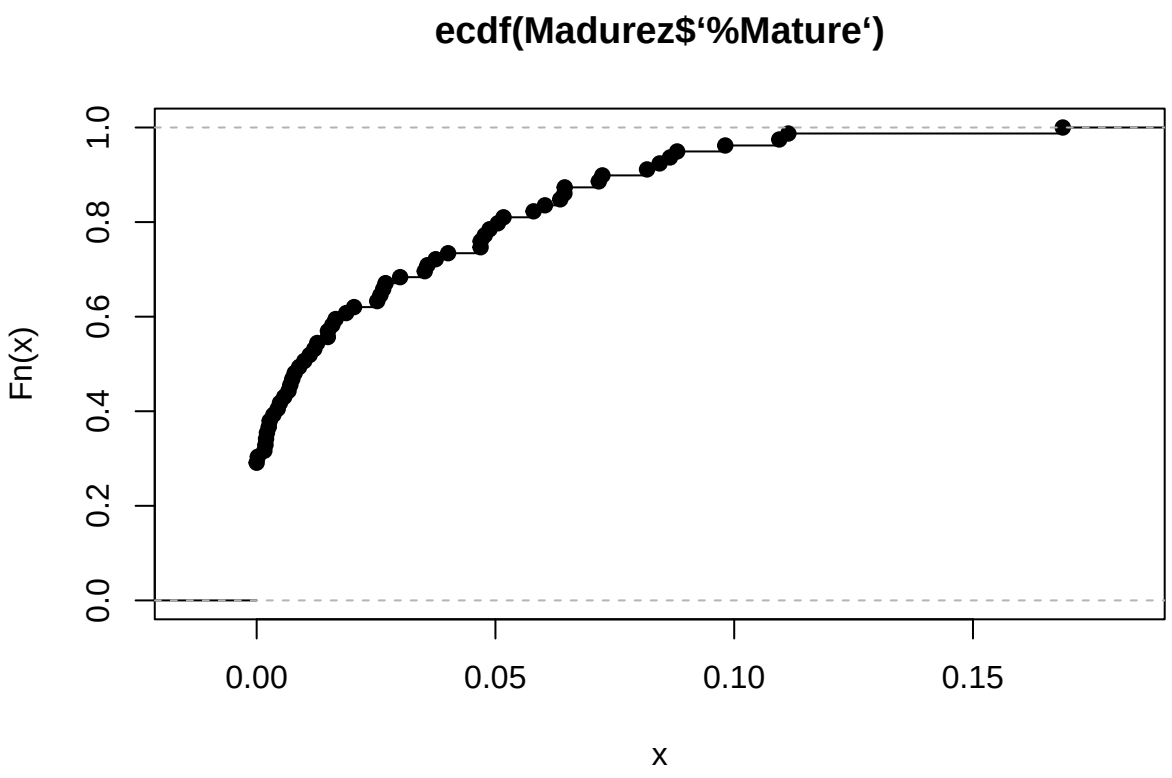
```
#hist(Madurez$Fotoperiodo,  main = "Fotoperiodo", col = "red")

Hist_Madurez<-ggplot(Madurez, aes(x = `%Mature`)) +geom_histogram(bins = 30, color = "black", fill="red

 P<-Madurez %>% select("Potencia") %>% ggplot(aes(x= Potencia))+
  geom_histogram(binwidth = 1, alpha=0.9, position = "identity", color="gold")+
  ylab("Frecuencia")+
  ggtitle("Histograma Madurez")

grid.arrange(Hist_Madurez, P, nrow =2)
```



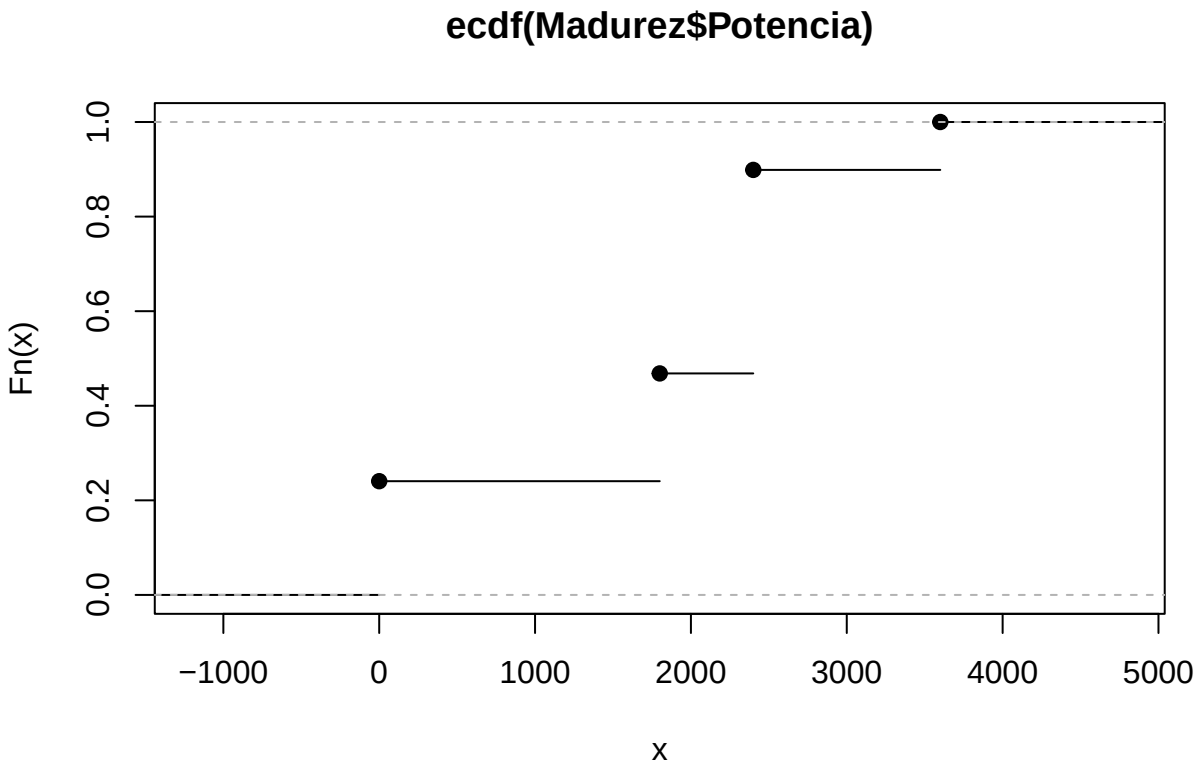

```
plot(density(Madurez$Potencia))
```



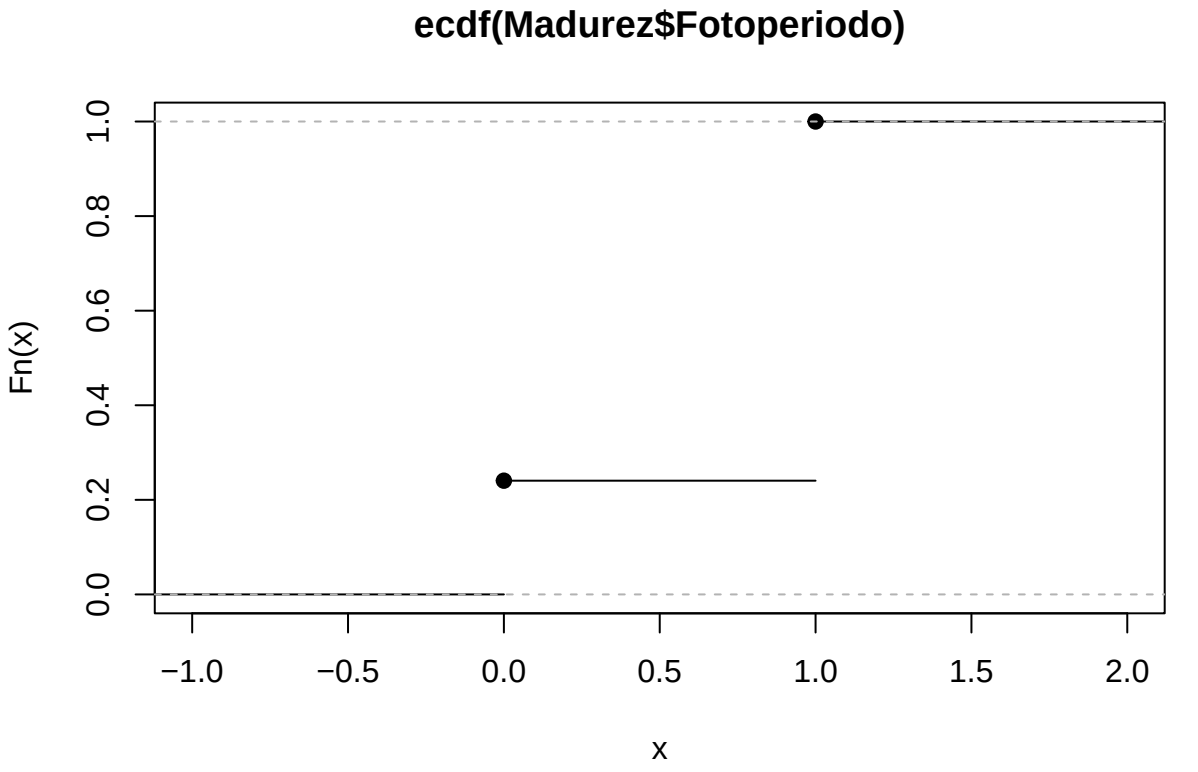

```
plot(ecdf(Madurez$`%Mature`))
```




```
plot(ecdf(Madurez$Potencia))
```

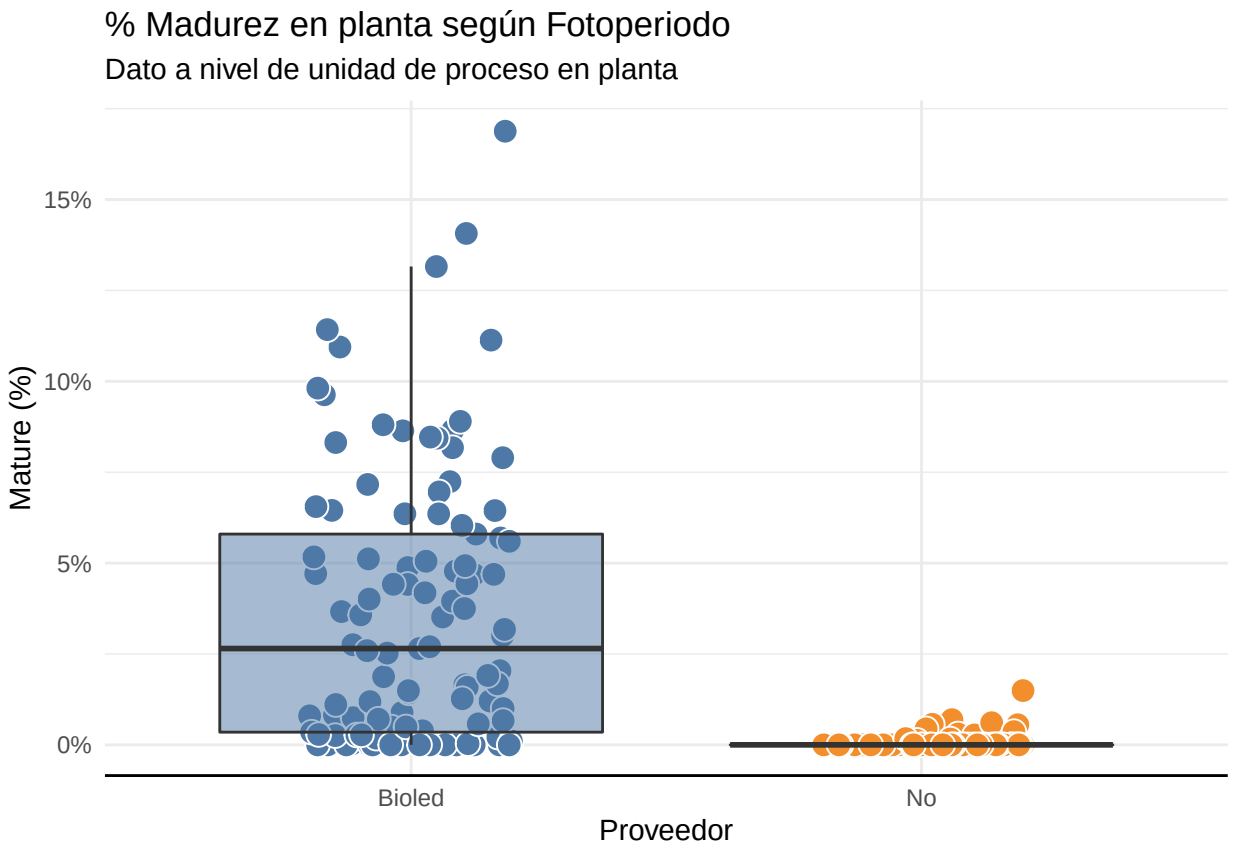



```
plot(ecdf(Madurez$Fotoperiodo))
```



Exploración por Proveedor

```
# Madurez ~ Proveedor
Madurez <- read_excel("Coho Season 2021-2022.xlsx", sheet = 2) %>%
  clean_names()
(Madurez <- Madurez %>%
   filter(!is.na(percent_mature)) %>%
   mutate(proveedor = fct_relevel(proveedor)) %>%
  ggplot(aes(proveedor, percent_mature, fill=proveedor)) +
   geom_jitter(shape=21, size=4, color="white", width = 0.2) +
   geom_boxplot(alpha=.5, outlier.color = "NA") +
   scale_fill_tableau() +
   theme_minimal() +
   scale_y_continuous(labels = percent) +
   labs(title="% Madurez en planta según Fotoperiodo",
        subtitle = "Dato a nivel de unidad de proceso en planta") +
   theme(axis.line.x = element_line(),legend.position='none')
)+
labs( x="Proveedor",
     y="Mature (%)")
```



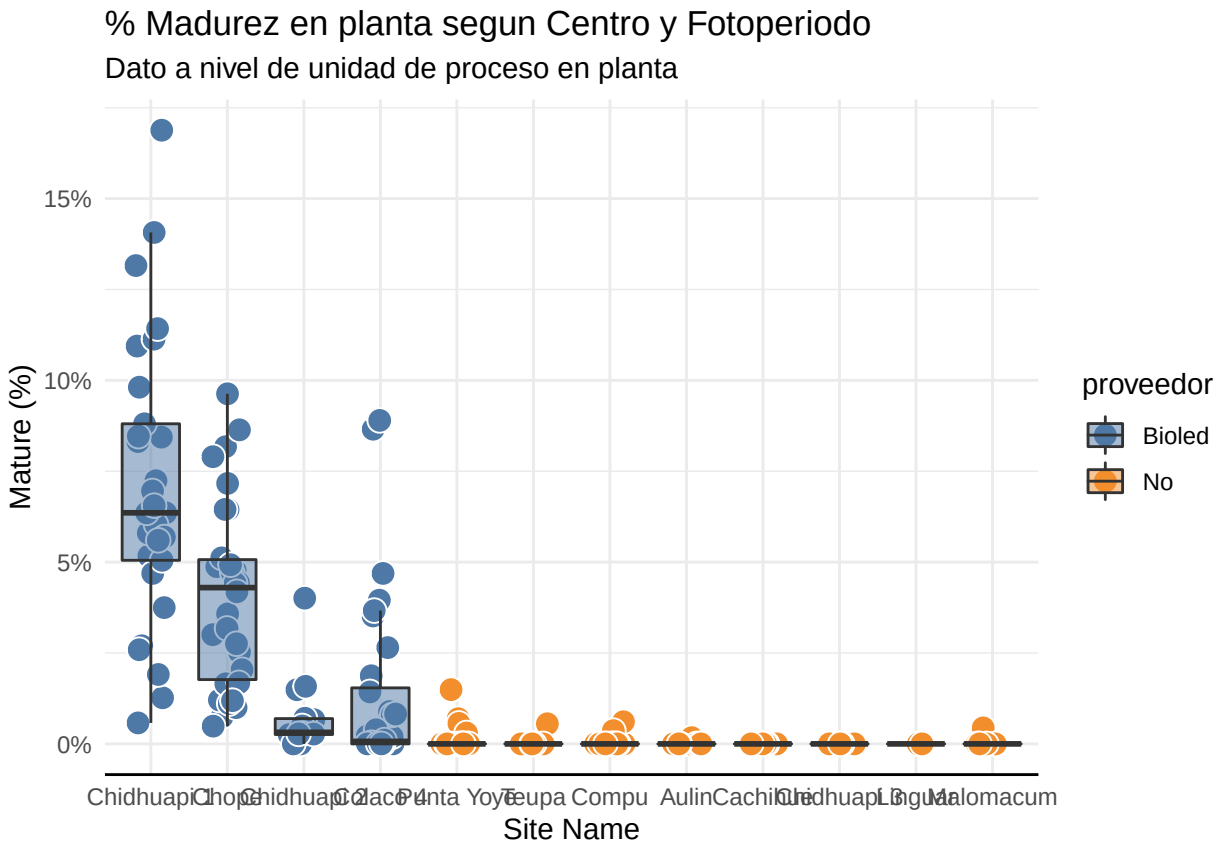
Exploración por Centro & Fotoperiodo

```
#Madurez  ~ Centro + Fotoperiodo
Madurez <- read_excel("Coho Season 2021-2022.xlsx", sheet = 2) %>%
  clean_names()
Madurez$site_name <- factor(Madurez$site_name, levels = c('Chidhuapi 1', 'Chope','Chidhuapi 2', 'Colaco
```

```
Madurez
```

```
## # A tibble: 220 x 11
##    jaula_individual site_name cage    percent_deformation percent_mature
##               <dbl> <fct>     <chr>                 <dbl>          <dbl>
##  1                1 Teupa     102                  0.0680        0
##  2                1 Teupa     103                  0.0451        0
##  3                1 Teupa     104                  0.0464        0
##  4                1 Teupa     107                  0.0692        0
##  5                0 Teupa     101/105              0.0117        0.00553
##  6                0 Teupa     101-108              0.0334        0
##  7                0 Teupa     102/107              0.0539        0
##  8                0 Teupa     102-103              0.0975        0
##  9                0 Teupa     103-108              0.0658        0
## 10                0 Teupa     104-106              0.0914        0
## # ... with 210 more rows, and 6 more variables: percent_desadaptado <dbl>,
## #   origen <chr>, tipo_orgien <chr>, fotoperiodo <dbl>, proveedor <chr>,
## #   potencia <dbl>
```

```
(Madurez <- Madurez %>%
   filter(!is.na(percent_mature)) %>%
   mutate(proveedor = fct_relevel(proveedor)) %>%
   ggplot(aes(site_name, percent_mature, fill=proveedor)) +
   geom_jitter(shape=21, size=4, color="white", width = 0.2) +
   geom_boxplot(alpha=.5, outlier.color = "NA") +
   scale_fill_tableau() +
   theme_minimal() +
   scale_y_continuous(labels = percent) +
   labs(title="% Madurez en planta segun Centro y Fotoperiodo",
        subtitle = "Dato a nivel de unidad de proceso en planta") +
   theme(axis.line.x = element_line())
)+
  labs( x="Site Name",
        y="Mature (%)")
```



Exploración por Madurez & Potencia

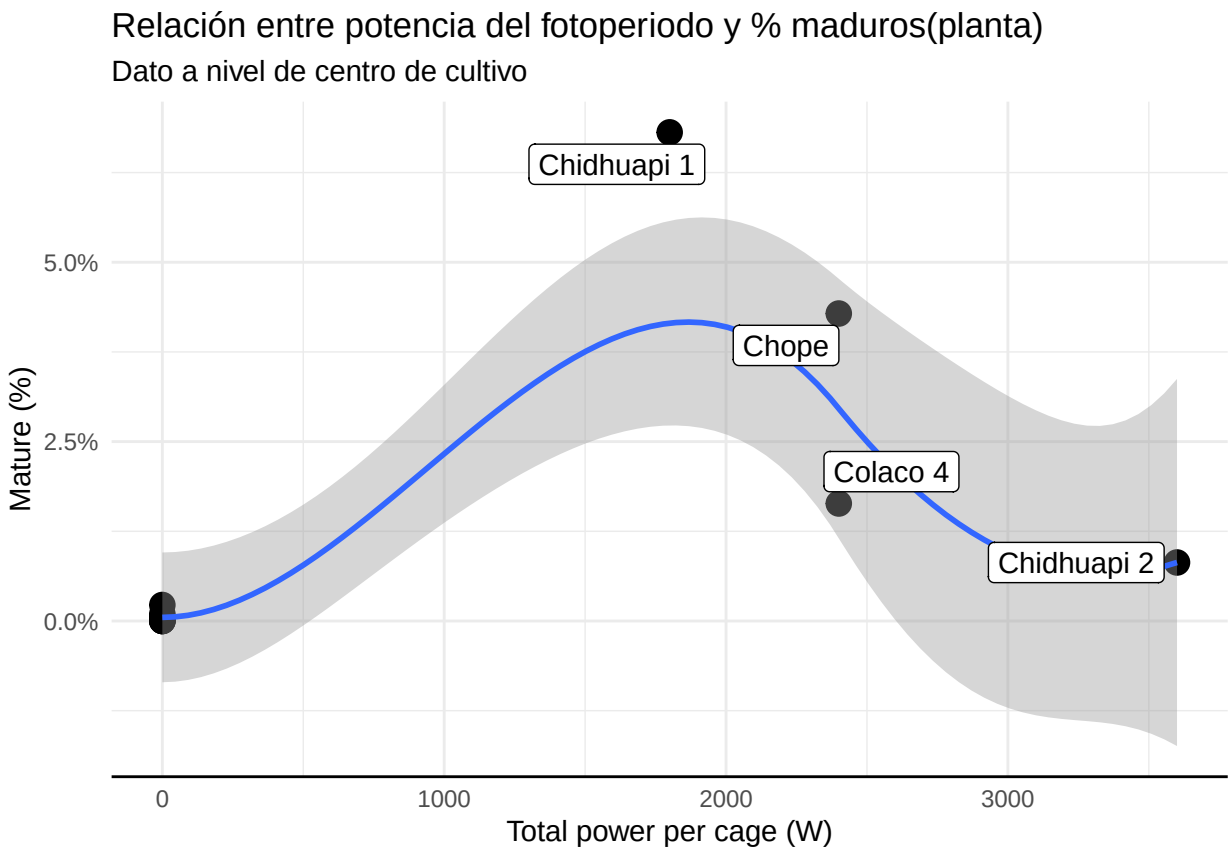
```
##Madurez ~ Potencia
datos_centro <- read_excel("Coho Season 2021-2022.xlsx", sheet = "Site detail", skip = 2) %>%
  clean_names()
#maduros-potencia (W)
(maduros_potencia_centro <- datos_centro %>%
   ggplot(aes(potencia,percent_maduro)) +
   geom_point(size=4) +
   geom_smooth() +
   theme_minimal() +
   #scale_x_continuous(labels = percent) +
   scale_y_continuous(labels = percent) +
   geom_label_repel(aes(label=site_name)) +
   labs(title="Relación entre potencia del fotoperiodo y % maduros(planta)",
        subtitle = "Dato a nivel de centro de cultivo") +
   theme(axis.line.x = element_line())
)+
labs( x="Total power per cage (W)",
        y="Mature (%)")#+
```

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

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at -18
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
```

```
## parametric, : neighborhood radius 1818

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 3.24e+06

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used at
## -18

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius 1818

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal condition
## number 0

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : There are other near
## singularities as well. 3.24e+06

## Warning: ggrepel: 8 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



```
#ylim(0,7)
```

Resume los datos usando tablas y estadística descriptiva

```
Ma<-table(Madurez$`%Mature`)
Fo<-table(Madurez$Fotoperiodo)
Po<-table(Madurez$Potencia)
mean(Madurez$`%Mature`)
```

```
## Warning in mean.default(Madurez$`%Mature`): argument is not numeric or logical:
## returning NA
```

```
## [1] NA
```

```
mean(Madurez$Fotoperiodo)
```

```
## Warning in mean.default(Madurez$Fotoperiodo): argument is not numeric or
## logical: returning NA
```

```
## [1] NA
```

```
mean(Madurez$Potencia)
```

```
## Warning in mean.default(Madurez$Potencia): argument is not numeric or logical:
## returning NA
```

```
## [1] NA
```

```
sd(Madurez$`%Mature`)
```

```
## [1] NA
```

```
sd(Madurez$Fotoperiodo)
```

```
## [1] NA
```

```
sd(Madurez$Potencia)
```

```
## [1] NA
```

Conclusiones

De acuerdo a la exploración de datos, se puede evidenciar que el proveedor fotoperiodo, presento maduración en los centros de cultivos donde se implentó la estrategia en comparación con los centros de cultivo donde no se encontraba implementada la estrategia de Fotoperiodo.

El Centro de Cultivo que presentó mayor madurez ($> 5\%$) correspondio al centro Chidhuapi 1

Enlace a proyecto Github: https://github.com/Ictiosapiens/Tarea_DiplomadoR