CLASE 06 - MANIPULACIÓN AVANZADA DE DATOS

Diplomado en Análisis de Datos y Modelamiento Predictivo con Aprendizaje Automático para la Acuicultura.

Dra. María Angélica Rueda Calderón & Dr. José Gallardo Matus

Pontificia Universidad Católica de Valparaíso

25 April 2023

PLAN DE LA CLASE

1.- Introducción

- Limpieza de datos.
- Funciones comunes para limpieza de datos.
- Funciones avanzadas para limpieza de datos.
- Gráficas en panel: una o más variables.

2). Práctica con R y Rstudio cloud.

- Realizar manipulación de datos con tidyr y dplyr.
- Realizar gráficas avanzadas con ggplot2.

TAREAS COMUNES DE LA LIMPIEZA DE DATOS:

- Identificar y tratar valores duplicados y atípicos.
- Tratar valores faltantes mediante la eliminación o la imputación.
- Normalizar/estandarizar las variables para que estén en una escala común.
- Anonimizar datos personales.
- Verificar la consistencia y la integridad de los datos.

IMPORTANCIA DE LA LIMPIEZA DE DATOS

¿Por qué es importante la limpieza de datos?

- Mejora la precisión y eficiencia del análisis.
- Mejora la toma de decisiones.
- Identifica patrones y tendencias ocultas en los datos.
- Cumplir con regulaciones y normativas de protección de datos.
- Mejora la estimación de parámetros estadísticos lo que reduce costos.

PROBLEMÁTICAS COMUNES DEL ANALISTA DE DATOS

- **Valores faltantes**: Eliminar observaciones (filas) o imputar valores faltantes.
- **Valores atípicos**: Identificarlos (Boxplot) y tratarlos adecuadamente (eliminar o corregir).
- **Datos inconsistentes**: diferentes fuentes, como errores de entrada de datos o diferencias en la forma en que se registran los datos.
- **Datos duplicados**: pueden surgir de la combinación de diferentes conjuntos de datos o errores en la entrada/codificación de los datos.
- **Datos desactualizados**: Actualizar regularmente para garantizar la precisión y la calidad de los datos.
- **Falta de documentación**: Realizar documentación detallada sobre el proceso de la limpieza y la manipulación de los datos.

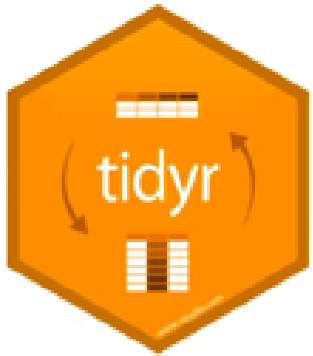
FUNCIONES DE R QUE PODEMOS USAR EN LA LIMPIEZA DE DATOS

- Verificar la estructura de los datos utilizando `str()`
- Obtener vista previa de los datos `head()`
- Verificar la consistencia y la integridad de los datos mediante la función `summary()`
- Identificar valores atípicos o extremos mediante gráfico **boxplot()**

FUNCIONES AVANZADAS DE R QUE PODEMOS USAR EN LA LIMPIEZA DE DATOS

- Identificar y tratar valores faltantes `is.na()`
- Eliminar filas de datos faltantes `na.omit()`
- Imputar valores faltantes `replace_na()`
- Identificar y tratar valores duplicados `duplicated()`
- Unificar datos duplicados ‘`distinct()`
- Normalizar/estandarizar los datos para ajustarlos a una escala común `scale()` o `normalize()`

PAQUETES CLAVE



PASOS DE LIMPIEZA DE DATOS

Base de datos de salmones con **252** observaciones (filas) y **6** variables (columnas).

- Revisar si hay datos faltantes (NA), duplicados o atípicos.

Sample	Ploidy	Family	Tank	Weight	Length
M100	Triploid	5	16	43	15.0
M100	Diploid	19	15	41	14.5
M1002	Diploid	9	16	NA	14.3
M1006	Diploid	9	16	38	14.6
M1010	Diploid	5	16	39	14.6
M1016	Diploid	19	16	29	13.2
M102	Diploid	19	15	39	14.3
M1022	Diploid	15	16	38	14.0
M1036	Diploid	1	16	33	13.5
M1041	Diploid	1	16	44	15.1

ESTRUCTURA DE LA BASE DE DATOS

- Verificar la estructura de los datos utilizando **`str()`**
- Transformar variables que están en formato ***chr*** a ***factor*** con **`as.factor()`**
- Transformar a variable númerica con **`as.numeric()`**

```
tibble [252 × 6] (S3: tbl_df/tbl/data.frame)
 $ Sample: Factor w/ 250 levels "M100","M1002",..: 1 1 2 3 4 5 6 7 8 9 ...
 $ Ploidy: Factor w/ 2 levels "Diploid","Triploid": 2 1 1 1 1 1 1 1 1 1 ...
 $ Family: Factor w/ 7 levels "1","11","15",..: 6 5 7 7 6 5 5 3 1 1 ...
 $ Tank  : Factor w/ 2 levels "15","16": 2 1 2 2 2 2 1 2 2 2 ...
 $ Weight: num [1:252] 43 41 NA 38 39 29 39 38 33 44 ...
 $ Length: num [1:252] 15 14.5 14.3 14.6 14.6 13.2 14.3 14 13.5 15.1 ...
```

IDENTIFICACIÓN DE NAs

- Revisar si hay datos faltantes (NA), atípicos, duplicados con `summary()`.

```
summary(salmon)
```

```
     Sample         Ploidy    Family  Tank         Weight           Length      
 M100   :  2   Diploid :233   1 :40   15:129   Min.   :  4.00   Min.   :-12.00  
 M1307  :  2   Triploid: 19   11:33   16:123   1st Qu.: 32.00   1st Qu.: 13.40  
 M1002  :  1                  15:42            Median : 37.00   Median : 14.00  
 M1006  :  1                  17:31            Mean   : 37.94   Mean   : 13.72  
 M1010  :  1                  19:42            3rd Qu.: 41.00   3rd Qu.: 14.60  
 M1016  :  1                  5 :27            Max.   :540.00   Max.   : 16.00  
 (Other):244                  9 :37            NA's   :2        NA's   :1       
```

ELIMINACIÓN DE NAs

- Dimensión de la base de datos con datos faltantes

```
## [1] 252   6
```

```
dim(salmon)
```

- Omitir/quitar datos faltantes `na.omit()`

```
salmon_new <- na.omit(salmon)
```

- Dimensión de la nueva base de datos sin datos faltantes

```
dim(salmon_new)
```

```
## [1] 250   6
```

REEMPLAZAR/IMPUTAR NAs

Reemplazar datos faltantes por la media, mediana, etc

`replace_na()`

```
salmon <- salmon %>% mutate(Weight =
replace_na(Weight,median(Weight, na.rm =
TRUE)),Length = replace_na(Length,median(Length,
na.rm = TRUE)))
```

Sample	Ploidy	Family	Tank	Weight	Length
M100	Triploid	5	16	43	15.0
M100	Diploid	19	15	41	14.5
M1002	Diploid	9	16	37	14.3
M1006	Diploid	9	16	38	14.6

IDENTIFICAR DATOS DUPLICADOS

Revisar si hay observaciones duplicadas con **`duplicated()`**

- La información en todas las columnas está duplicada

```
dups_all <- salmon %>% filter(duplicated(.))
```

Sample	Ploidy	Family	Tank	Weight	Length
M1307	Diploid	19	16	48	15.5

- Las observaciones están duplicadas para el mismo individuo

```
dups_id <- salmon %>% filter(duplicated(Sample))
```

Sample	Ploidy	Family	Tank	Weight	Length
M100	Diploid	19	15	41	14.5
M1307	Diploid	19	16	48	15.5

UNIFICAR DATOS DUPLICADOS

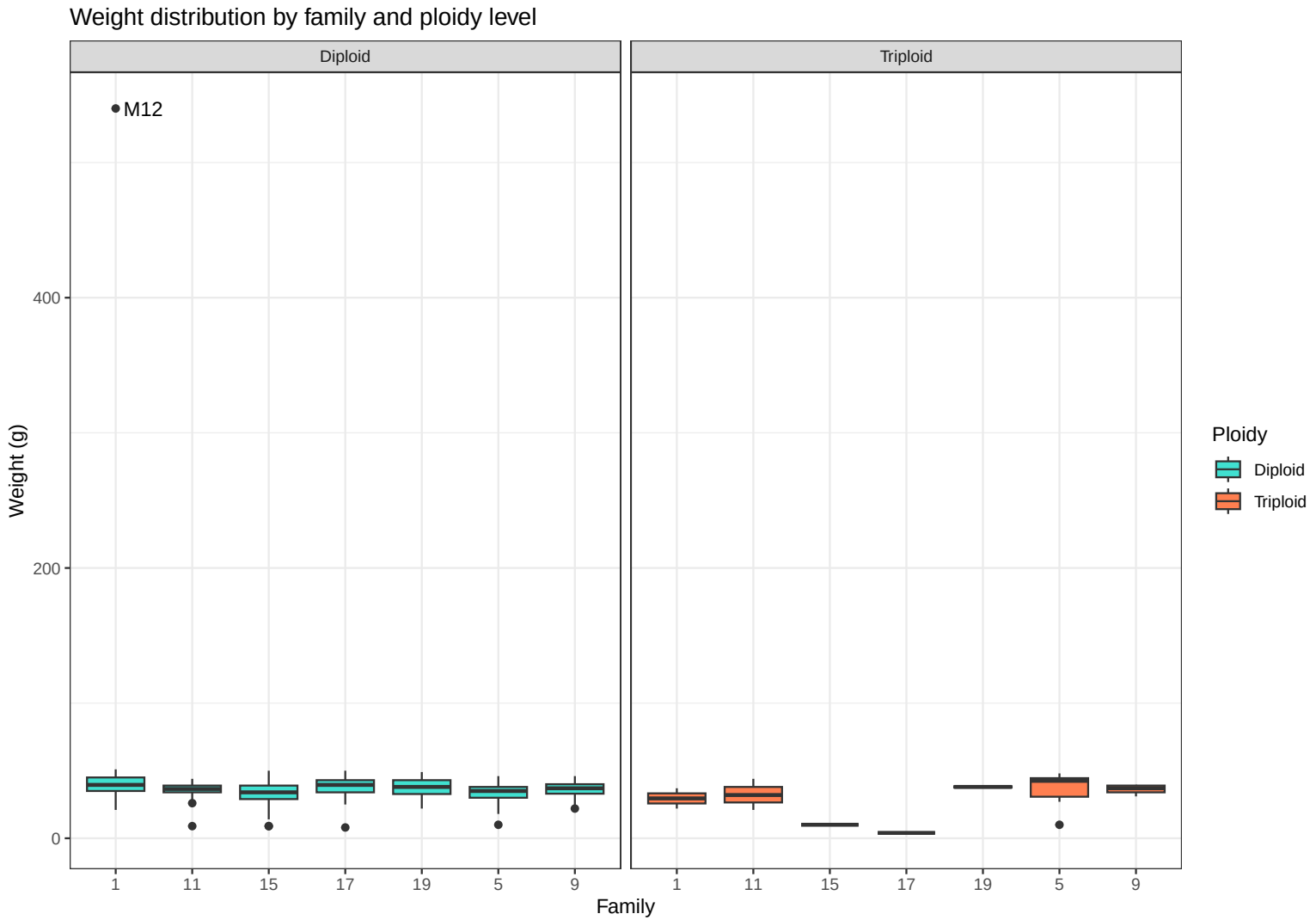
Corregir o unificar datos duplicados `distinct()`

```
salmon_unified <- salmon %>% distinct(Sample,
.keep_all = TRUE)
```

Sample	Ploidy	Family	Tank	Weight	Length
M100	Triploid	5	16	43	15.0
M1002	Diploid	9	16	37	14.3
M1006	Diploid	9	16	38	14.6
M1010	Diploid	5	16	39	14.6

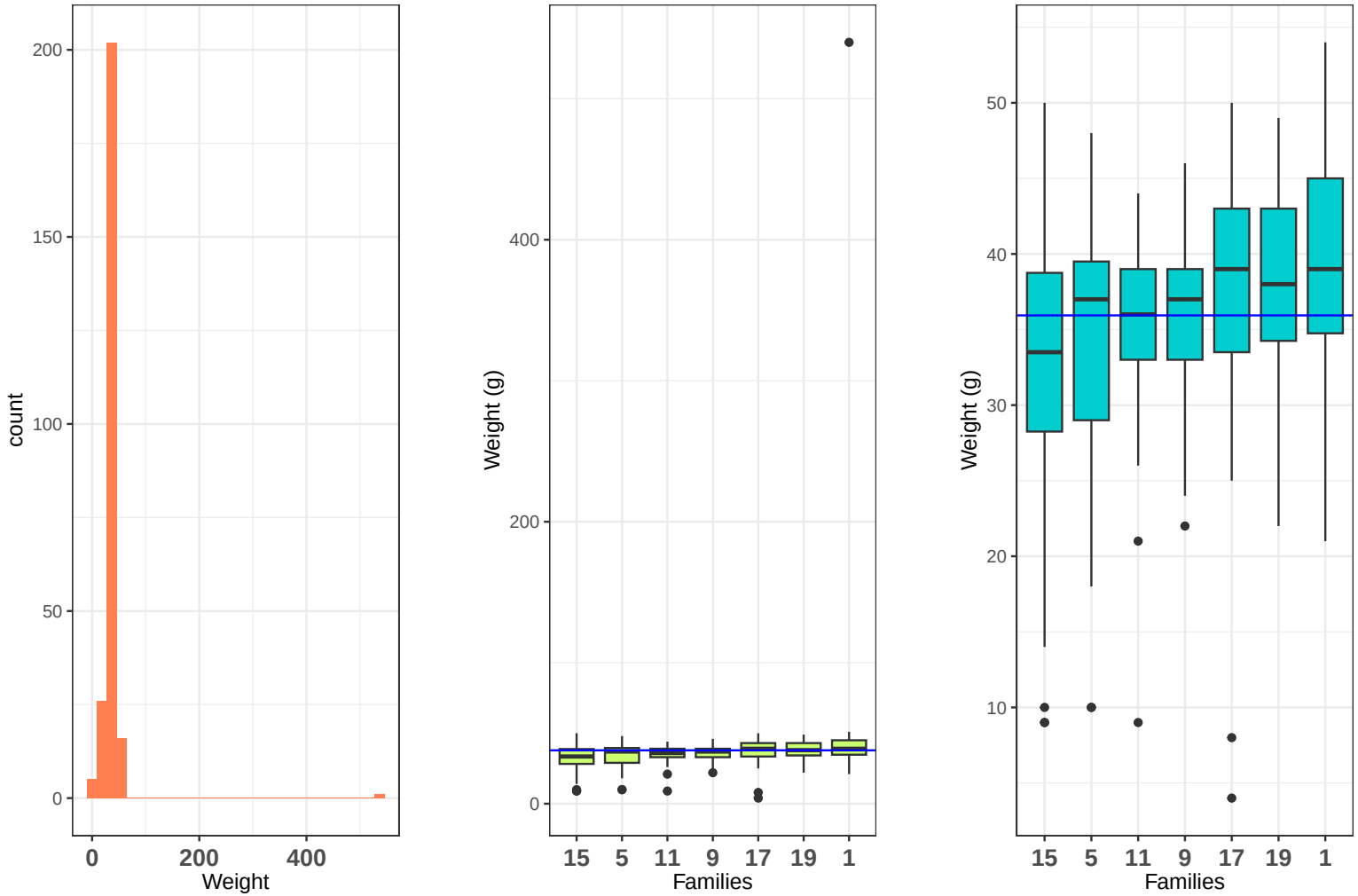
VISUALIZAR POR NIVEL DE UN FACTOR

`facet_wrap()`: Permite dividir una gráfica en paneles o subgráficos basados en una o varias variables categóricas.



GRÁFICOS EN PANEL

`grid.arrange(p, q, r, ncol = 3)` : se utiliza para combinar varios gráficos de ggplot en una sola figura.



RESUMEN DE LA CLASE

- Identificamos valores atípicos, duplicados y faltantes.
- Usamos funciones avanzadas para unificar valores duplicados, imputación de datos faltantes y modificación de valores atípicos.
- Utilizamos tuberías o pipe **%>%**.
- Hicimos gráficos ggplot2 usando funciones de manipulación avanzada de datos.