Limpieza y formateo de la base de datos Arequipa

1. Se utilizaron las siguientes hojas de Excel: Desembarque, biologico y biometrico

2. Se formatearon los nombres en mayúsculas de todos los rótulos de todas las columnas

3. Si formatearon las columnas Matricula y ... , borrando los caracteres alfanuméricos y guiones

# Library -----------------------------------------------------------------

Llamamos los paquetes a utilizar a través de la función ***library***

library(readxl)

library(openxlsx)

# Data -------------------------------------------------------------------

Ubicamos las bases de datos a utilizar para luego leerlas con la función ***read\_xlsx*** , donde se especifica el número de la hoja de cálculo con la que se trabajará; además se eliminan las filas vacías superiores a los rótulos donde estas existan

# View(base\_biom) --------------------------------------------------------------

Se procede a convertir en mayúsculas los rótulos de cada columna a través de la función ***toupper*** para cada hoja de cálculo a trabajar

names(base\_cap) <- toupper(names(base\_cap))

names(base\_biol) <- toupper(names(base\_biol))

names(base\_biom) <- toupper(names(base\_biom))

Para las columnas MATRIC y MATRICULA , de las data frame *Captura* y *Biometrico* respectivamente, se eliminarón todos los caracteres alfanuméricos y guiones

base\_cap$MATRIC <- gsub("[^[:alnum:]-]", "", base\_cap$MATRIC)

base\_biom$MATRICULA <- gsub("[^[:alnum:]-]", "", base\_biom$MATRICULA)

Para la columna mes de la data frame base\_cap se colocó un cero adelante del dato todos aquellos que presentaron una cifra

base\_cap$MES <- ifelse(nchar(base\_cap$MES) == 1, paste0("0", base\_cap$MES), base\_cap$MES)

base\_cap$DIA <- ifelse(nchar(base\_cap$DIA) == 1, paste0("0", base\_cap$DIA), base\_cap$DIA)

Se creó la columna KEY dentro de la data frame base\_cap, la cual constó de la concatenación de las columnas AÑO, MES, DIA y MATRIC , y los que se unieron por guiones

base\_cap$KEY <- paste(base\_cap$AÑO, base\_cap$MES, base\_cap$DIA,

base\_cap$MATRIC, sep = "-")

Para la columna FECHA de la data frame base\_biom se eliminó los caracteres “UTC”

base\_biom$FECHA <- gsub(pattern = " UTC", replacement = "", base\_biom$FECHA)

Se creó la columna KEY dentro de la data frame base\_biom, la cual constó de la concatenación de las columnas FECHA y MATRIC , cuya separación fue por guiones

base\_biom$KEY <- paste(base\_biom$FECHA, base\_biom$MATRICULA, sep = "-")

Se eliminaron los espacios en blancos al inicio y fin de los datos de las columnas KEY para la data frame tanto de base\_cap como para base\_biom

base\_cap$KEY <- trimws(base\_cap$KEY)

base\_biom$KEY <- trimws(base\_biom$KEY)

Se utilizó la función ***merge*** para combinar dos data frame, base\_biom y un conjunto de columnas de la data frame base\_cap, tomando como valor de coincidencia la columna KEY; además se indicó una unión izquierda para que todas las filas del data frame base\_biom se mantengan aunque no haya coincidencia con las filas seleccionadas de la data frame base\_cap, lo cual se rellenaria con NA

base\_biom <- merge(base\_biom, base\_cap[, c("KEY", "FABRICA", "CBOD(M3)")], by = " KEY", all.x = TRUE)

# Variables ---------------------------------------------------------------

Se creo un vector, l\_sp, con elementos que incrementan de 0.5 en 0.5 que inica en 5 y finaliza en 20 ; además se creó un segundo vector , ls\_sp, donde los datos incrementan de 1 en 1 y van desde el 5 hasta el 1

l\_sp <- seq(5, 20, .5)

ls\_sp <- seq(5, 70, 1)

Se obtuvo el valor absoluto de las columnas LONG y LAT para luego multiplicar por -1 con la finalidad de obtener coordenadas negativas corresponidientes a Perú

base\_biom[,"LONG"] <- -1\*abs(base\_biom[, "LONG"])

base\_biom[,"LAT"] <- -1\*abs(base\_biom[, "LAT"])

La columna ESPECIE se rellenó con "Anchoveta, anchoveta peruana, peladilla" ya que toda la base era sobre anchoveta

base\_biom[,"ESPECIE"] <- "Anchoveta, anchoveta peruana, peladilla"

Se renombró la columan CAPTURA (t) por CAPTURA(t)

base\_biom[,"CAPTURA(t)"] <- base\_biom[, "CAPTURA (t)"]

Se creó la columna región , la cual fue rellenada con "AREQUIPA"

base\_biom[,"REGION"] <- "AREQUIPA"

Para las columnas FABRICA, EMBARCACION, TIPO DE FLOTA, PUERTO y AREA no se realizó ninguna modificación

La columna CBOD(M3) fue renombrada por CB

base\_biom[,"CB"] <- base\_biom[, "CBOD(M3)"]

La columna FECHA fue renombrada por DIA

base\_biom[,"DIA"] <- base\_biom[, "FECHA"]

Se creó una columna con nombre Otras y se rellenó con NA porque no habían datos

base\_biom[,"Otras"] <- NA

Para completar las columnas faltantes , longitudes de otras especies (ls\_sp), se creó una matriz con el número de filas del tamaño de la data frame base\_biom y con el número de columnas de ls\_sp; previamente se creo un vector con datos secuenciales desde el 5 hasta el 70 que incrementaban de 1 en 1 (ls\_sp). Luego se convirtió en data frame (df), para un mejor manejo de los datos. Las columnas de df se denomidaron como LS\_5 , LS\_6, LS\_7, así sucesivamente hasta LS\_70

df <- as.data.frame.matrix(matrix(nrow = nrow(base\_biom), ncol = length(ls\_sp)))

names(df) <- paste("LS", ls\_sp, sep = "\_")

Seguidamente se creó el vector con todas las columnas previamente tratadas

base\_biom <- base\_biom[, c("LON", "LAT", "ESPECIE", "CAPTURA(t)", "REGION",

"FABRICA", "EMBARCACION", "MATRICULA", "CB", "TIPO DE FLOTA",

"PUERTO", "AREA", "DIA", as.character(l\_sp), "Otras")]

Finalmente se crea un nuevo data frame a través de la función ***cbind***, la cual combinará las columnas de base\_biom y df, de tal manera que tendríamos la base de datos de Arequipa adaptada a una de IMARSIS ; esta se guarda como csv para poder visualizarse en un excel.

base\_biom <- cbind(base\_biom, df)

write.csv(base\_biom, "cin/BD\_DESEMBARQUE\_AREQU…DESCARGA.csv", row.names = F)