Bitacora 2

Bitácora 2 Proyecto Grupal

Integrantes:

- Sebastián Calderón Segura C21517
- Sebastián Miranda Ramírez C4H274
- Josué Gustavo Rodríguez Aguilar C4J023
- Kevin David Calderon Martínez C4D511
- Santiago Paniagua Chavarría C4I249

La presente bitácora documenta el proceso de análisis, depuración y exploración de los datos correspondientes a los delitos registrados en Costa Rica entre 2019 y julio de 2025, obtenidos del portal de datos abiertos del Poder Judicial. A lo largo del informe se describe el tratamiento de las bases de datos originales, la unificación de variables, la detección de valores faltantes y atípicos, así como la aplicación de técnicas estadísticas para garantizar la coherencia y calidad del conjunto de datos. El objetivo principal es obtener una base limpia y estructurada que permita interpretar con mayor precisión las tendencias delictivas en el país.

1 Características de los datos

Se utilizaron las bases de datos de crímenes registrados entre 2019 y julio de 2025, disponibles en el portal de datos del Organismo de Investigación Judicial (OIJ) (2025)..

En este sentido, estas bases venían separadas por año y presentaban algunas inconsistencias entre ellas. Por ejemplo, entre 2019 y 2023 la variable del sexo se denominaba “género”, mientras que a partir de 2024 pasó a llamarse “sexo”. Asimismo, la columna de fecha aparecía en formato *char* en algunos años y en formato *Date* en otros.

Para resolver esto, se unificaron los datos, combinando todos los archivos en una única base que incluye todos los crímenes registrados en el periodo 2019–julio 2025.

De esta forma, resultó una base de datos con las siguientes características:

- Tamaño: 345065 x 12.
- Va de 2019-01-01 a 2025-07-31.
- Todas las variables son cualitativas: delito, subdelito, fecha, hora, victima, subvictima, edad, sexo, nacionalidad, provincia, canton, distrito.

Población de estudio

La población de estudio corresponde a todos los eventos delictivos ocurridos en Costa Rica de 2019 a julio de 2025.

Muestra observada

La muestra son las observaciones registradas de delitos en Costa Rica del 01/01/2019 a las 00:00:00 al 31/07/25 a las 23:59:59, almacenadas en la base del Poder Judicial.

Unidad estadística

Corresponde a cada observación de delito junto con sus atributos: delito, subdelito, fecha, hora, victima, subvictima, edad, sexo, nacionalidad, provincia, canton, distrito

Variables de estudio

Variable	Tipo	Descripción
fecha	Fecha	Día del evento
hora	Categórica	Tramos horarios de 3 horas
delito	Categórica	Categoría principal del hecho
subdelito	Categórica	Subtipo de delito
victima	Categórica	Tipo de víctima (pueden ser objetos como VIVIENDA o VEHÍCULO).
subvictima	Categórica	Subtipo de víctima
sexo	Categórica	Sexo de la víctima.
edad	Categórica	Edad de la víctima.
nacionalidad	Categórica	Nacionalidad de la víctima.
provincia	Categórica	Provincia donde ocurrió el hecho
canton	Categórica	Cantón donde ocurrió el hecho

Variable	Tipo	Descripción
distrito	Categórica	Distrito donde ocurrió el hecho

Es importante recalcar que las bases de datos traen columnas correspondientes a persona (edad, sexo, nacionalidad) incluso cuando victima es VEHICULO, VIVIENDA, EDIFICACION, etc; lo cual se puede deber a valores por defecto de captura. Para un correcto análisis se coloca NA en dichas posiciones. Aqui se muestra un ejemplo de como quedarían los datos aplicando estos cambios:

```
datos %>%
  select(victima, subvictima, edad, sexo, nacionalidad) %>%
  filter(victima != "PERSONA") %>%
  slice_head(n = 10)
```

```
# A tibble: 10 x 5
   victima                 subvictima edad      sexo      nacionalidad
   <chr>                   <chr>      <chr>     <chr>     <chr>
 1 VIVIENDA [NO APLICA]    NO APLICA  NO APLICA NO APLICA NO APLICA
 2 EDIFICACION [NO APLICA] NO APLICA  NO APLICA NO APLICA NO APLICA
 3 EDIFICACION [NO APLICA] NO APLICA  NO APLICA NO APLICA NO APLICA
 4 EDIFICACION [NO APLICA] NO APLICA  NO APLICA NO APLICA NO APLICA
 5 VEHICULO [NO APLICA]    NO APLICA  NO APLICA NO APLICA NO APLICA
 6 VIVIENDA [NO APLICA]    NO APLICA  NO APLICA NO APLICA NO APLICA
 7 VIVIENDA [NO APLICA]    NO APLICA  NO APLICA NO APLICA NO APLICA
 8 VIVIENDA [NO APLICA]    NO APLICA  NO APLICA NO APLICA NO APLICA
 9 VEHICULO [NO APLICA]    NO APLICA  NO APLICA NO APLICA NO APLICA
10 EDIFICACION [NO APLICA] NO APLICA  NO APLICA NO APLICA NO APLICA
```

2 Muestra Parcial de la Base de Datos

Se agregan las primeras 5 filas de nuestra base de datos:

```
slice_head(datos, n = 5)
```

```
# A tibble: 5 x 12
  delito subdelito  fecha      hora  victima subvictima edad  sexo  nacionalidad
  <chr>  <chr>      <date>     <chr> <chr>   <chr>      <chr> <chr> <chr>
1 ASALTO ARMA BLAN~ 2019-01-01 18:0~ PERSONA PEATON [P~ MAYO~ HOMB~ COSTA RICA
```

```
2 ASALTO ARMA BLAN~ 2019-01-01 06:0~ PERSONA PEATON [P~ MAYO~ HOMB~ COSTA RICA
3 ASALTO ARMA BLAN~ 2019-01-01 18:0~ PERSONA OTRO O IN~ MAYO~ HOMB~ COSTA RICA
4 ASALTO ARMA BLAN~ 2019-01-01 03:0~ PERSONA PEATON [P~ MAYO~ MUJER COSTA RICA
5 ASALTO ARMA BLAN~ 2019-01-01 12:0~ VIVIEN~ NO APLICA  NO A~ NO A~ NO APLICA
# i 3 more variables: provincia <chr>, canton <chr>, distrito <chr>
```

3 Muestras de las Variables Cualitativas

Uno de los principales retos encontrados en esta base de datos, es que todas las variables de la misma son categóricas. Es decir, no hay ninguna variable cuantitativa. Incluso la variable de edad, se divide en pocas categorias (NO APLICA, ADULTO MAYOR, MAYOR DE EDAD, MENOR DE EDAD, DESCONOCIDO)

Entonces, para afrontar este reto, se realiza una distribucion de frecuencias como la que se ha mostrado anteriormente. De esta manera, se puede analizar, retomando el caso de la edad por ejemplo, la frecuencia de cada categoria. Pero, se analiza cada variable de acuerdo al orden de la base de datos:

```
resumen_delito <-  read_csv(here("data/processed/resumen_delito.csv"), show_col_types = FALSE
resumen_delito
```

```
# A tibble: 10 x 2
   delito                            frecuencia_de_delitos
   <chr>                                             <dbl>
 1 HURTO                                             79231
 2 ASALTO                                            67374
 3 ROBO                                              55178
 4 DELITOS CONTRA LA PROPIEDAD                       45854
 5 ROBO DE VEHICULO                                  23138
 6 TACHA DE VEHICULO                                 19683
 7 ESTAFAS Y OTRAS DEFRAUDACIONES                    18302
 8 DELITOS CONTRA LA VIDA                             7971
 9 OTROS DELITOS CONTRA LA PROPIEDAD                  5421
10 DELITOS INFORMATICOS                               4903
```

Estos son los 10 delitos que más frecuencia ha tenido en el pais en los ultimos seis años, aunque no se debe descartar que hay mas de 30 tipos de delitos reportados en los ultimos seis años. Lideran el hurto, asalto, robo, los delitos contra la propiedad y el vehículo como los cinco delitos más frecuentes. Por lo que se puede entender que hay una tendencia hacia los delitos reportados que implican el robo de un bien.

```
resumen_victima <-  read_csv(here("data/processed/resumen_victima.csv"), show_col_types = FA
resumen_victima
```

```
# A tibble: 5 x 2
  victima                 frecuencia_de_victimas
  <chr>                                    <dbl>
1 PERSONA                                 159090
2 VEHICULO [NO APLICA]                     68422
3 VIVIENDA [NO APLICA]                     61221
4 EDIFICACION [NO APLICA]                  41144
5 OTROS [NO APLICA]                        15188
```

Con respecto a las víctimas, aproximadamente la mitad de las víctimas de los delitos reportados son personas. Note quel total de víctimas de personas se relaciona con los primeros tres delitos más reportados (Hurto, asalto y robo). Seguido de los vehículos (Robo de vehículos), la vivienda y edificaciones (Delitos contra la propiedad) y los otros. Para efectos de esta investigación, el foco de atención se centra en las víctimas que son personas.

```
resumen_edad <-  read_csv(here("data/processed/resumen_edad.csv"), show_col_types = FALSE)
resumen_edad
```

```
# A tibble: 5 x 2
  edad          frecuencia_rango_de_edades
  <chr>                              <dbl>
1 NO APLICA                         185975
2 MAYOR DE EDAD                     132656
3 MENOR DE EDAD                      10076
4 ADULTO MAYOR                        9273
5 DESCONOCIDO                         7085
```

De acuerdo a las frecuencias, se interpretan las mismas como rangos de edad. Más de la mitad de las víctimas no son personas (Como se observa en el cuadro de víctimas), por lo que, las personas mayores de edad son las que más reportan delitos que ocurren en contra de sí mismas. También se puede observar que bajo los mayores de edad, se encuentran los menores; por lo que han habido más de 10 000 víctimas menores de edad en los últimos 6 años. En menor frecuencia los adultos mayores, y por último, más de 7000 casos con edad no registrada.

```
resumen_sexo <-  read_csv(here("data/processed/resumen_sexo.csv"), show_col_types = FALSE)
resumen_sexo
```

```
# A tibble: 4 x 2
  sexo        frecuencia_de_sexo
  <chr>                    <dbl>
1 NO APLICA               185975
2 DESCONOCIDO              66142
3 HOMBRE                   58178
4 MUJER                    34770
```

Análogamente con la edad, hay exactamente 185975 casos donde las víctimas no son personas. Por ende, solo hay registro de más de 58 mil hombres víctimas de delitos, y más de 34 mil víctimas mujeres. Se desconoce el sexo de 66000 casos aproximadamente.

```
resumen_nacionalidad <-  read_csv(here("data/processed/resumen_nacionalidad.csv"), show_col_t
resumen_nacionalidad
```

```
# A tibble: 10 x 2
   nacionalidad   frecuencia_de_nacionalidades
   <chr>                                 <dbl>
 1 NO APLICA                            185975
 2 COSTA RICA                           126959
 3 NICARAGUA                             17179
 4 DESCONOCIDO                            4710
 5 ESTADOS UNIDOS                         1899
 6 VENEZUELA                               821
 7 ALEMANIA                                723
 8 COLOMBIA                                612
 9 FRANCIA                                 575
10 CANADA                                  530
```

Obviando los casos donde las víctimas no son personas, aproximadamente el 37% de las víctimas son de nacionalidad costarricense, seguidamente del 5% de las víctimas, que son nicaraguenses, el 1,36% son casos no registrados y el 0,55% de víctimas han sido estadounidenses.

```
resumen_provincia <-  read_csv(here("data/processed/resumen_provincia.csv"), show_col_types =
resumen_provincia
```

```
# A tibble: 8 x 2
  provincia   frecuencia_de_provincias
  <chr>                          <dbl>
1 SAN JOSE                      129730
```

```
2 ALAJUELA                    54118
3 PUNTARENAS                 41217
4 LIMON                      32249
5 HEREDIA                    30854
6 GUANACASTE                 29394
7 CARTAGO                    27474
8 DESCONOCIDO                   29
```

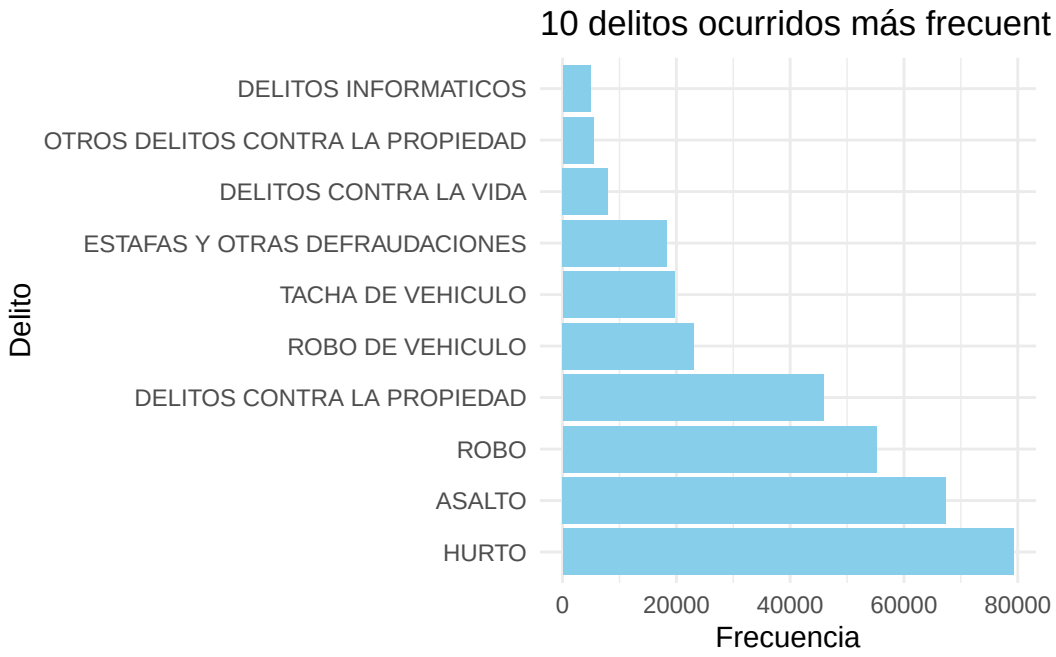
Acá se puede observar la frecuencia de delitos cometidos de acuerdo a la provincia. Siendo la capital la provincia donde más se reportan delitos, seguido de Alajuela, con una diferencia de reporte de delitos de 75612 casos, una variación muy alta, a diferencia de las siguientes provincias. Por último, Cartago, la provincia donde menos se reportan delitos, y 29 casos de los que no se tiene información.

4 Gráficos de las Variables

Se ha hecho un grafico de la distribucion de cada tabla de frecuencia. Aca se muestran las mas relevantes:

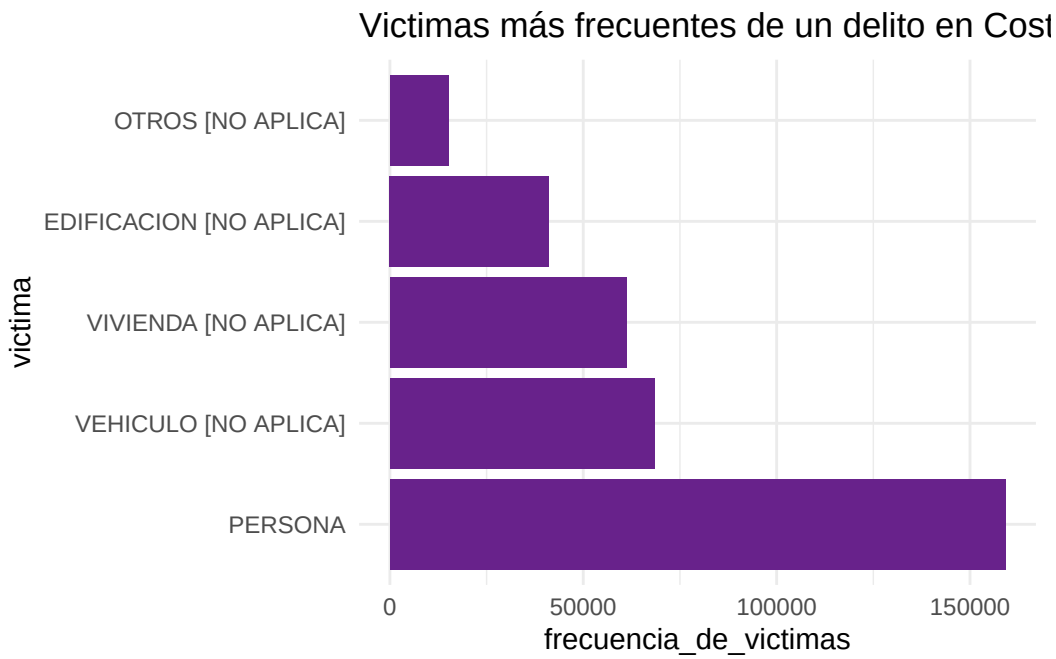
Variable delito:

```
grafico_delito <- resumen_delito %>%
  arrange(desc(frecuencia_de_delitos)) %>%
  mutate(delito = factor(delito, levels = delito)) %>%
  ggplot(aes(x =delito, y = frecuencia_de_delitos, fil = delito)) +
  geom_bar(stat = "identity", fill = "skyblue", show.legend = FALSE) +
  coord_flip() +
  labs(
    title = "10 delitos ocurridos más frecuentres en Costa Rica entre 2019 y 2025",
    x = "Delito",
    y = "Frecuencia"
  ) +
  theme_minimal()
grafico_delito
```



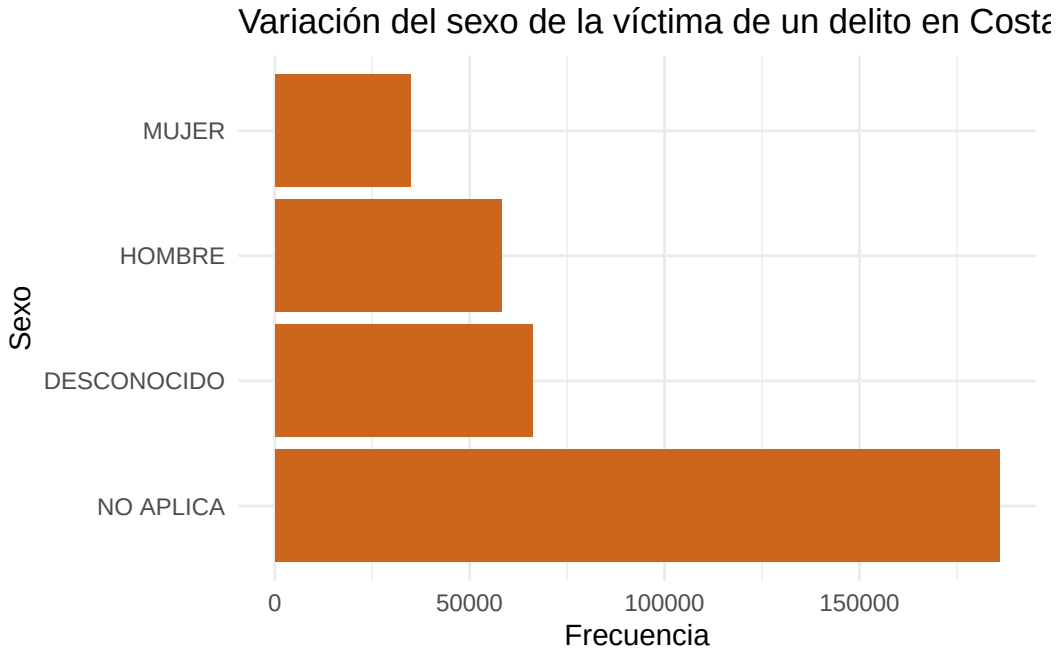
Variable víctima:

```
grafico_victima <- resumen_victima %>%
  mutate(victima = factor(victima, levels = victima)) %>%
  ggplot(aes(x = victima, y = frecuencia_de_victimas)) +
  geom_bar(stat = "identity", fill = "darkorchid4") +
  coord_flip()+
  labs(title = "Victimas más frecuentes de un delito en Costa Rica entre 2019 y 2025") +
  theme_minimal()
#Guardamos
ggsave(
  filename = here("info", "graphics", "grafico_victima.png"),
  plot = grafico_victima,
  width = 15, height = 10, dpi = 200
)
grafico_victima
```



Variable sexo:

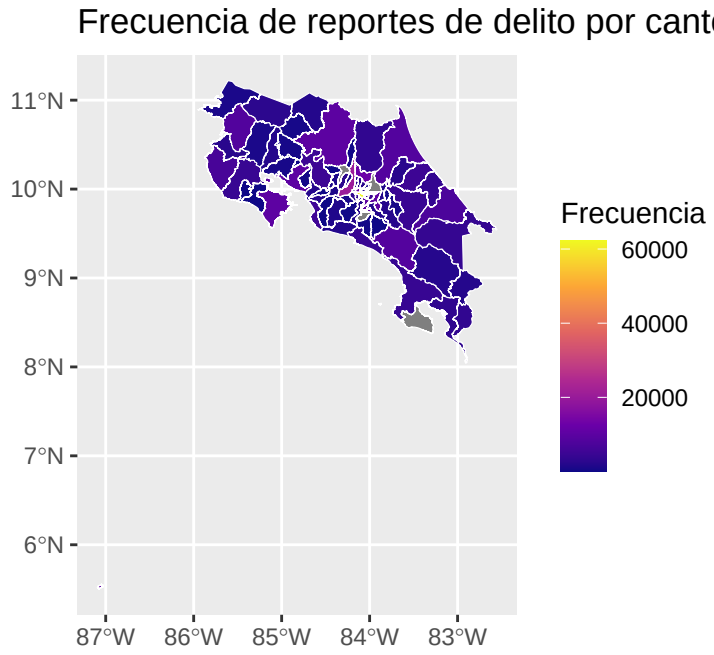
```
grafico_sexo <- resumen_sexo %>%
  mutate(sexo = factor(sexo, levels = sexo)) %>%
  ggplot(aes(x = sexo, y = frecuencia_de_sexo)) +
  geom_bar(stat = "identity", fill = "chocolate3" ) +
  coord_flip() +
  labs(title = "Variación del sexo de la víctima de un delito en Costa Rica entre 2019 y 2025
       x = "Sexo",
       y = "Frecuencia") +
  theme_minimal()
#Guardamos
ggsave(
  filename = here("info", "graphics", "grafico_sexo.png"),
  plot = grafico_sexo,
  width = 15, height = 10, dpi = 200
)
grafico_sexo
```



Para el caso de la variable cantón, se utiliza una base de datos geoespacial (Instituto Nacional de Estadística y Censos, 2024). Esta divide poligonalmente el mapa de Costa Rica de acuerdo a los cantones:

```
resumen_canton <- datos %>%
  count(canton, name = "frecuencia_de_cantones") %>%
  arrange(desc(frecuencia_de_cantones))

resumen_canton <- resumen_canton %>%
  filter(canton != "DESCONOCIDO")

cantones <- st_read("../data_raw/shapefiles/unidad_geoestadistica_cantonal_ugec_2024.shp",qu

cantones <- cantones  %>%
  mutate(nomb_ugec = str_to_upper(nomb_ugec),
         nomb_ugec = str_replace_all(nomb_ugec, c(
           "Á" = "A", "É" = "E", "Í" = "I", "Ó" = "O", "Ú" = "U", "Ñ" = "N"
         )))

cantones <- cantones %>%
  left_join(resumen_canton, by = c("nomb_ugec"= "canton"))

grafico_cantones <- ggplot(cantones) +
```

```
  geom_sf(aes(fill = frecuencia_de_cantones), color = "white") +
  scale_fill_viridis_c(option = "plasma") +
  labs(title = "Frecuencia de reportes de delito por canton en Costa Rica entre 2019 y 2025"
grafico_cantones
```



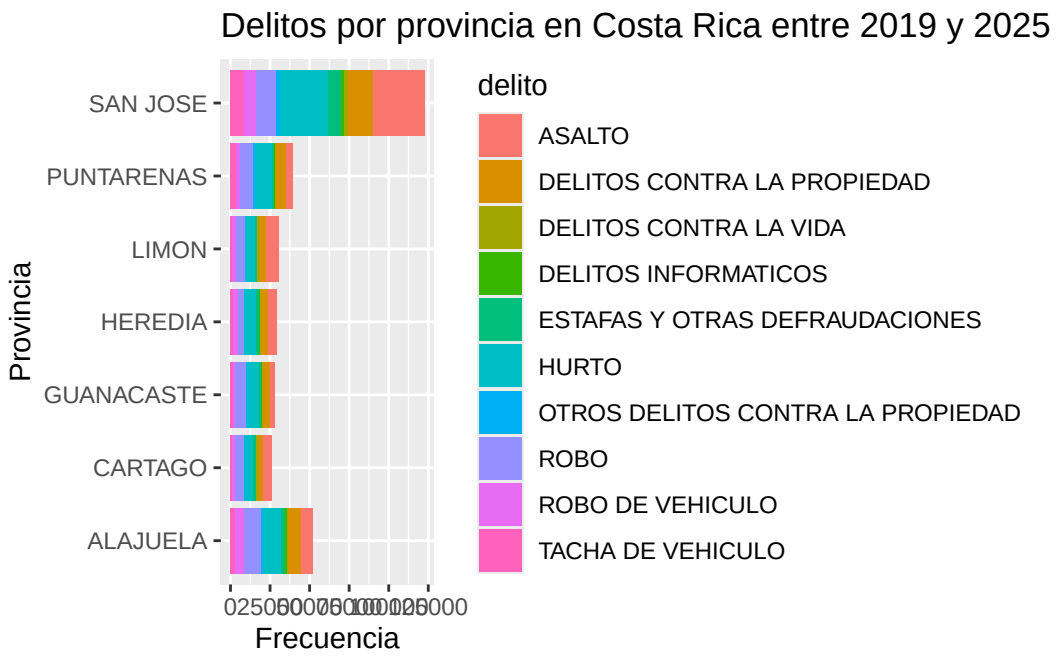
5 Comparación de Variables

Al ser variables categoricas, se debe de tener coherencia a la hora de comparar variables.

En este caso, se compara primero, la variable de provincia con la de delito, es decir, contabilizar los distintos tipos de delitos que ocurren en Costa Rica en funcion a su provincia:

```
ruta.entrada <- here("data", "processed", "Estadísticas Policiales 2019 a Julio 2025.csv")
datos <- read_csv(ruta.entrada, show_col_types = FALSE)

#Df enfocaddo en relacionar los delitos por provincia
filtro <- datos %>%
  filter(!provincia %in% c("DESCONOCIDO", "NO APLICA"),
         delito != "DESCONOCIDO")

delitos_mas_frecuentes <-  filtro %>%
  group_by(delito) %>%
```

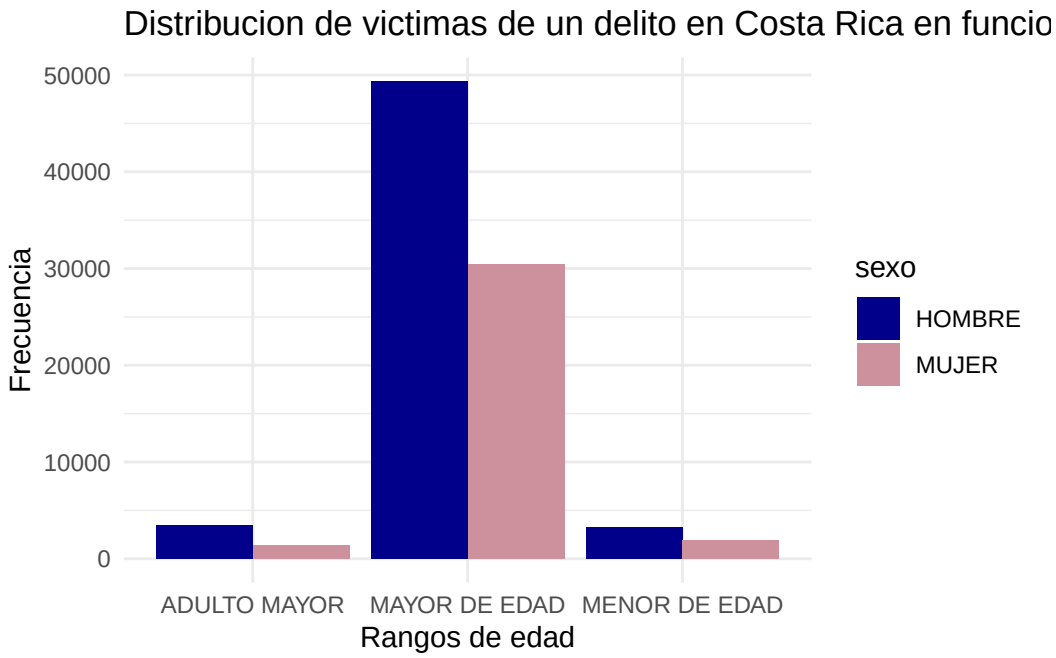
```r
  summarise(frecuencia.delito = n(), .groups = "drop") %>%
  arrange(desc(frecuencia.delito)) %>%
  slice_head(n = 10) %>%
  pull(delito)

delito_provincia <- filtro %>%
  filter(delito %in% delitos_mas_frecuentes) %>%
  group_by(provincia, delito) %>%
  summarise(frecuencia = n(), .groups = "drop") %>%
  mutate(provincia = factor(provincia), delito = factor(delito))

#Relacion entre fechas y provincias
grafico_delito_provincia <- delito_provincia %>%
  ggplot(aes(x = provincia, y = frecuencia, fill = delito)) +
  geom_bar(stat ="identity") +
  coord_flip() +
  labs(title = "Delitos por provincia en Costa Rica entre 2019 y 2025",
       x = "Provincia",
       y ="Frecuencia")

grafico_delito_provincia
```



Ahora, la relacion entre la variable sexo y edad. Es decir, la distribucion de los rangos de edad

de acuerdo al sexo que han reportado un delito en Costa Rica entre 2019 y 2025:

```
resumen_sexo_edad <- datos %>%
  filter(!edad %in% c("DESCONOCIDO", "NO APLICA"),
         !sexo %in% c("DESCONOCIDO", "NO APLICA")) %>%
  group_by(sexo, edad) %>%
  summarise(frecuencia = n(), .groups = "drop")

#Grafico
grafico_sexo_edad <- resumen_sexo_edad %>%
  ggplot(aes(x = edad, y = frecuencia, fill = sexo)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_manual(values = c("HOMBRE" = "blue4", "MUJER" = "pink3")) +
  labs(title = "Distribucion de victimas de un delito en Costa Rica en funcion del sexo y la
       x = "Rangos de edad",
       y = "Frecuencia") +
  theme_minimal()
grafico_sexo_edad
```



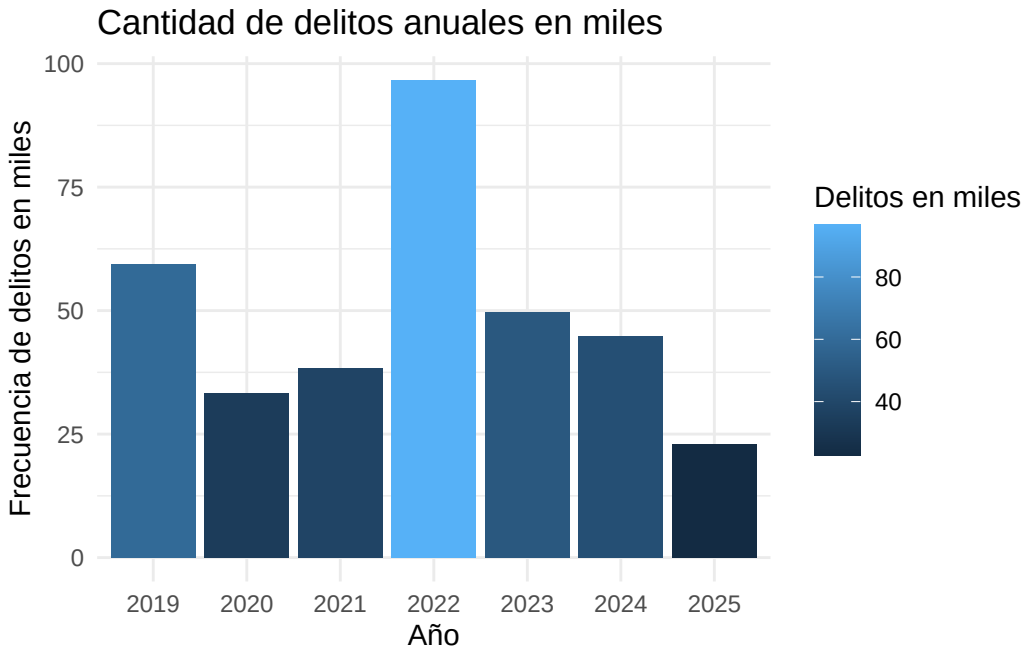
6 Gráfico de Variable Fecha

De manera representativa para este caso se realizara un grafico de la variable categorica fecha, pues se considera relevante para el analsis critico en el desarrollo del proyecto conocer la trayectoria de los delitos a travez del tiempo, ademas no se considetan otras varibales pues es reiterativo por el trato de las variables cualitativas:

```
#Ahora bien tenemos cargada la base de datos del proyecto, primero verificamos los tipos de v
str(datos)
```

```
spc_tbl_ [345,065 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ delito      : chr [1:345065] "ASALTO" "ASALTO" "ASALTO" "ASALTO" ...
 $ subdelito   : chr [1:345065] "ARMA BLANCA" "ARMA BLANCA" "ARMA BLANCA" "ARMA BLANCA" ...
 $ fecha       : Date[1:345065], format: "2019-01-01" "2019-01-01" ...
 $ hora        : chr [1:345065] "18:00:00 - 20:59:59" "06:00:00 - 08:59:59" "18:00:00 - 20:59
 $ victima     : chr [1:345065] "PERSONA" "PERSONA" "PERSONA" "PERSONA" ...
 $ subvictima  : chr [1:345065] "PEATON [PERSONA]" "PEATON [PERSONA]" "OTRO O INDETERMINADO
 $ edad        : chr [1:345065] "MAYOR DE EDAD" "MAYOR DE EDAD" "MAYOR DE EDAD" "MAYOR DE EDA
 $ sexo        : chr [1:345065] "HOMBRE" "HOMBRE" "HOMBRE" "MUJER" ...
 $ nacionalidad: chr [1:345065] "COSTA RICA" "COSTA RICA" "COSTA RICA" "COSTA RICA" ...
 $ provincia   : chr [1:345065] "GUANACASTE" "SAN JOSE" "PUNTARENAS" "SAN JOSE" ...
 $ canton      : chr [1:345065] "SANTA CRUZ" "ESCAZU" "PUNTARENAS" "SAN JOSE" ...
 $ distrito    : chr [1:345065] "TAMARINDO" "ESCAZU" "EL ROBLE" "HOSPITAL" ...
 - attr(*, "spec")=
  .. cols(
  ..   delito = col_character(),
  ..   subdelito = col_character(),
  ..   fecha = col_date(format = ""),
  ..   hora = col_character(),
  ..   victima = col_character(),
  ..   subvictima = col_character(),
  ..   edad = col_character(),
  ..   sexo = col_character(),
  ..   nacionalidad = col_character(),
  ..   provincia = col_character(),
  ..   canton = col_character(),
  ..   distrito = col_character()
  .. )
 - attr(*, "problems")=<externalptr>
```

```r
#Note que de hecho la totalidad de las variables utilizadas son categoricas por lo que se pr

grafico_fecha_frecuencia <- datos %>%
  mutate(anio = year(fecha)) %>%
  group_by(anio) %>%
  summarise(frecuencia = n()) %>%
  mutate(frecuencia_miles = frecuencia / 1000) %>% #Antes el grafico se veia incosistente co
  ggplot(aes(x = factor(anio),y = frecuencia_miles,  fill = frecuencia_miles)) +
  geom_col() +
  theme_minimal() +
  labs(
    x = "Año",
    y = "Frecuencia de delitos en miles",
    title = "Cantidad de delitos anuales en miles",
    fill = "Delitos en miles"
  )
grafico_fecha_frecuencia
```



Tome en cuenta que el proyecto se compone de variables categoricas, que en su mayoria fueron consideradas en los puntos 4 y 5. Por lo tanto para este insiso unicamente se tomo en cuenta las fechas.

7 Outliers y Valores Vacíos

Para identificar los valores faltantes, es útil convertir los valores que aparecen como NO APLICA o DESCONOCIDO en verdaderos valores faltantes (NA) solo cuando el valor no tiene sentido en función del tipo de víctima. Para esto se puede modelar una función que detecta NA en cada archivo y genera un reporte de cuántos valores faltan.

```
reporte_faltantes_df <- function(df, dataset_name = "", etiquetas = c("DESCONOCIDO", "NO APL
  total_rows   <- nrow(df)
  na_reales    <- sapply(df, function(x) sum(is.na(x)))
  na_etiquetas <- sapply(df, function(x) if (is.character(x)) sum(x %in% etiquetas) else 0L)
  data.frame(
    dataset      = dataset_name,
    columna      = names(df),
    na_reales    = as.integer(na_reales),
    na_etiqueta  = as.integer(na_etiquetas),
    na_total     = as.integer(na_reales + na_etiquetas),
    pct_faltante = round(100 * (na_reales + na_etiquetas) / max(total_rows, 1L), 2),
    row.names = NULL
  )
}

# Reportes (uno por cada resumen)
reporte_faltantes <- rbind(
  reporte_faltantes_df(resumen_canton,       "resumen_canton"),
  reporte_faltantes_df(resumen_delito,       "resumen_delito"),
  reporte_faltantes_df(resumen_edad,         "resumen_edad"),
  reporte_faltantes_df(resumen_nacionalidad, "resumen_nacionalidad"),
  reporte_faltantes_df(resumen_provincia,    "resumen_provincia"),
  reporte_faltantes_df(resumen_sexo,         "resumen_sexo"),
  reporte_faltantes_df(resumen_victima,      "resumen_victima")
)

# Mostrar el reporte completo
reporte_faltantes
```

```
                dataset                      columna na_reales na_etiqueta
1        resumen_canton                       canton         0           0
2        resumen_canton      frecuencia_de_cantones         0           0
3        resumen_delito                       delito         0           0
4        resumen_delito       frecuencia_de_delitos         0           0
5          resumen_edad                         edad         0           2
```

```
6          resumen_edad   frecuencia_rango_de_edades         0           0
7  resumen_nacionalidad                 nacionalidad         0           2
8  resumen_nacionalidad frecuencia_de_nacionalidades         0           0
9     resumen_provincia                     provincia         0           1
10    resumen_provincia     frecuencia_de_provincias         0           0
11         resumen_sexo                         sexo         0           2
12         resumen_sexo           frecuencia_de_sexo         0           0
13      resumen_victima                      victima         0           0
14      resumen_victima       frecuencia_de_victimas         0           0
   na_total pct_faltante
1         0          0.0
2         0          0.0
3         0          0.0
4         0          0.0
5         2         40.0
6         0          0.0
7         2         20.0
8         0          0.0
9         1         12.5
10        0          0.0
11        2         50.0
12        0          0.0
13        0          0.0
14        0          0.0
```

Ahora, para identificar outliers se va a utilizar el método IQR (Interquartile Range o Rango Intercuartílico) la cual es una técnica estadística utilizada para identificar valores atípicos en un conjunto de datos. Un valor se considera outlier si se encuentra fuera de los límites: $<$ Q1 $-$ 1.5 $\times$ IQR o $>$ Q3 $+$ 1.5 $\times$ IQR. Donde IQR=Q3-Q1. Este método es robusto porque no se ve afectado por la asimetría o valores extremos aislados. En este proyecto se utilizará el criterio del 1.5 $\times$ IQR, que es el mismo que emplean los gráficos de caja (boxplots) para detectar valores inusualmente altos o bajos en las frecuencias de delitos.

```r
# Función que detecta outliers por regla IQR
outliers_iqr <- function(df, id_col, val_col, dataset){
  v  <- df[[val_col]]
  q1 <- quantile(v, 0.25, na.rm = TRUE)
  q3 <- quantile(v, 0.75, na.rm = TRUE)
  iqr <- q3 - q1
  lim_inf <- q1 - 1.5 * iqr
  lim_sup <- q3 + 1.5 * iqr
```

```r
  out <- df[v < lim_inf | v > lim_sup, c(id_col, val_col), drop = FALSE]
  names(out) <- c("id", "valor")
  mutate(out, dataset = dataset, .before = 1)
}


# Aplicación a cada resumen
o_canton    <- outliers_iqr(resumen_canton,       "canton",       "frecuencia_de_cantones",
o_delito    <- outliers_iqr(resumen_delito,       "delito",       "frecuencia_de_delitos",
o_edad      <- outliers_iqr(resumen_edad,         "edad",         "frecuencia_rango_de_edades
o_nacional  <- outliers_iqr(resumen_nacionalidad, "nacionalidad", "frecuencia_de_nacionalidad
o_provincia <- outliers_iqr(resumen_provincia,    "provincia",    "frecuencia_de_provincias"
o_sexo      <- outliers_iqr(resumen_sexo,         "sexo",         "frecuencia_de_sexo",
o_victima   <- outliers_iqr(resumen_victima,      "victima",      "frecuencia_de_victimas",

outliers_consolidados <- bind_rows(
  o_canton, o_delito, o_edad, o_nacional, o_provincia, o_sexo, o_victima
)

outliers_consolidados
```

```
# A tibble: 13 x 3
   dataset              id            valor
   <chr>                <chr>         <dbl>
 1 resumen_canton       SAN JOSE      62111
 2 resumen_canton       ALAJUELA      20569
 3 resumen_canton       HEREDIA       12748
 4 resumen_canton       CARTAGO       10497
 5 resumen_canton       PUNTARENAS    10374
 6 resumen_canton       SAN CARLOS    10105
 7 resumen_canton       DESAMPARADOS   9999
 8 resumen_nacionalidad NO APLICA    185975
 9 resumen_nacionalidad COSTA RICA   126959
10 resumen_provincia    SAN JOSE     129730
11 resumen_provincia    DESCONOCIDO      29
12 resumen_sexo         NO APLICA    185975
13 resumen_victima      PERSONA      159090
```

8 Técnicas de Tratamiento para Valores Vacíos y Outliers

Se identificaron anteriormente la presencia de valores vacíos o NA en las bases de datos. Para esto, se aplica una estrategia distinta para diferentes tipos de variables, con el objetivo de obtener una base de datos más limpia y sin cosas innecesarias, el cual se muestra a continuación:

```
# Demográficas: quitar NO APLICA / DESCONOCIDO (para análisis sobre personas)
resumen_edad_clean <- resumen_edad %>% filter(!edad %in% lab_na)
resumen_sexo_clean <- resumen_sexo %>% filter(!sexo %in% lab_na)
resumen_nacionalidad_clean <- resumen_nacionalidad %>% filter(!nacionalidad %in% lab_na)

# Territoriales: quitar DESCONOCIDO
resumen_provincia_clean <- resumen_provincia %>% filter(!provincia %in% "DESCONOCIDO")
resumen_canton_clean    <- resumen_canton %>% filter(!canton %in% "DESCONOCIDO")

# Delito/Víctima: conservar todo
resumen_delito_clean  <- resumen_delito
resumen_victima_clean <- resumen_victima
```

En el punto anterior se observaron muchos outliers, los cual se debe corregir. La winsorización es una técnica estadística utilizada para reducir el efecto de valores extremos sin eliminarlos del conjunto de datos. Consiste en reemplazar los valores que superan los límites definidos (ya sea por percentiles o por el rango intercuartílico) por el valor más cercano dentro de esos límites. En este proyecto se aplicará una winsorización basada en el método IQR ($1.5 \times$ IQR), con el fin de suavizar la influencia de los cantones o delitos con frecuencias extremadamente altas, como San José o Hurto, los cuales, aunque representan fenómenos reales, podrían distorsionar la visualización y el análisis estadístico. De esta forma se conserva la información esencial sin perder la coherencia general de la distribución.

```
# Función winsorizar por IQR (1.5 x IQR)
winsorizar_iqr <- function(df, col_valor){
  v   <- df[[col_valor]]
  q1  <- quantile(v, 0.25, na.rm = TRUE)
  q3  <- quantile(v, 0.75, na.rm = TRUE)
  iqr <- q3 - q1
  lim_inf <- q1 - 1.5 * iqr
  lim_sup <- q3 + 1.5 * iqr
  df %>%
    mutate(
      !!paste0(col_valor, "_winsor") := pmin(pmax(.data[[col_valor]], lim_inf), lim_sup),
      !!paste0(col_valor, "_log")    := log10(.data[[col_valor]] + 1)
```

```
    )
}

# Winsorización por IQR
resumen_canton_w  <- winsorizar_iqr(resumen_canton_clean,       "frecuencia_de_cantones")
resumen_delito_w  <- winsorizar_iqr(resumen_delito_clean,        "frecuencia_de_delitos")
resumen_edad_w    <- winsorizar_iqr(resumen_edad_clean,          "frecuencia_rango_de_edades"
resumen_nac_w     <- winsorizar_iqr(resumen_nacionalidad_clean,  "frecuencia_de_nacionalidade
resumen_prov_w    <- winsorizar_iqr(resumen_provincia_clean,     "frecuencia_de_provincias")
resumen_sexo_w    <- winsorizar_iqr(resumen_sexo_clean,          "frecuencia_de_sexo")
resumen_victima_w <- winsorizar_iqr(resumen_victima_clean,       "frecuencia_de_victimas")

#vista rapida
head(resumen_canton_w)
```

```
# A tibble: 6 x 4
  canton    frecuencia_de_cantones frecuencia_de_canton~1 frecuencia_de_canton~2
  <chr>                      <int>                  <dbl>                  <dbl>
1 SAN JOSE                   62111                  9518.                   4.79
2 ALAJUELA                   20569                  9518.                   4.31
3 HEREDIA                    12748                  9518.                   4.11
4 CARTAGO                    10497                  9518.                   4.02
5 PUNTAREN~                  10374                  9518.                   4.02
6 SAN CARL~                  10105                  9518.                   4.00
# i abbreviated names: 1: frecuencia_de_cantones_winsor,
#   2: frecuencia_de_cantones_log
```

```
head(resumen_delito_w)
```

```
# A tibble: 6 x 4
  delito     frecuencia_de_delitos frecuencia_de_delito~1 frecuencia_de_delito~2
  <chr>                      <dbl>                  <dbl>                  <dbl>
1 HURTO                      79231                  79231                   4.90
2 ASALTO                     67374                  67374                   4.83
3 ROBO                       55178                  55178                   4.74
4 DELITOS C~                 45854                  45854                   4.66
5 ROBO DE V~                 23138                  23138                   4.36
6 TACHA DE ~                 19683                  19683                   4.29
# i abbreviated names: 1: frecuencia_de_delitos_winsor,
#   2: frecuencia_de_delitos_log
```

```r
head(resumen_prov_w)
```

```
# A tibble: 6 x 4
  provincia frecuencia_de_provin~1 frecuencia_de_provin~2 frecuencia_de_provin~3
  <chr>                      <dbl>                   <dbl>                   <dbl>
1 SAN JOSE                  129730                  73983.                   5.11
2 ALAJUELA                   54118                  54118                    4.73
3 PUNTAREN~                  41217                  41217                    4.62
4 LIMON                      32249                  32249                    4.51
5 HEREDIA                    30854                  30854                    4.49
6 GUANACAS~                  29394                  29394                    4.47
# i abbreviated names: 1: frecuencia_de_provincias,
#   2: frecuencia_de_provincias_winsor, 3: frecuencia_de_provincias_log
```

Tras la limpieza y análisis de las bases de datos, se logró identificar patrones relevantes y reducir el impacto de valores atípicos mediante la aplicación de la técnica de winsorización. Esto permitió mantener la integridad de la información sin distorsionar las distribuciones originales. En general, la bitácora refleja un proceso metodológico que asegura la confiabilidad de los datos para futuros análisis estadísticos y geoespaciales sobre la criminalidad en Costa Rica.

Bibliografia

Organismo de Investigación Judicial (OIJ). (2025). *Datos Abiertos*. Organismo de Investigación Judicial. https://sitiooij.poder-judicial.go.cr/index.php/ayuda/servicios-policiales/servicios-a-organizaciones/indice-de-transparencia-del-sector-publico-costarricense/datos-abiertos

ggplot. (2020). *Cartesian coordinates with x and y flipped*. Gggplot2. https://ggplot2.tidyverse.org/reference/co

Milton. (2019). *Paleta de Colores en R*. SCRIBD. https://es.scribd.com/document/432904420/Paleta-de-Colores-en-R

QGIS. (2023). *Mapscaping*. Read Write Shapefiles In R. https://mapscaping.com/read-write-shapefiles-in-r/

Mapas de Coropletas en R. (s. f.). Rcharts. https://r-charts.com/es/espacial/mapa-coropletas/

Instituto Nacional de Estadística y Censos. (2024). *Geoportal Estadístico del INEC*. INEC. https://geoportal.inec.go.cr/