Índice

1 Tema

1.1 Expresión del gen TP53 en el tejido mamario de pacientes y de personas en riesgo de cáncer de pecho: análisis por grupos etarios

El gen TP53, ampliamente reconocido en el campo médico por su desempeño en la supresión tumoral, pertenece al conjunto de genes que de acuerdo con López et al. (2001), están implicados en diversos procesos de división celular, dentro de ellos la regulación de la expresión génica, el control del ciclo celular, la programación de la muerte celular y la estabilidad del genoma, es decir, la capacidad de la célula para mantener su material genético sin modificaciones a lo largo del tiempo. El gen mencionado inicialmente tiene múltiples funciones, puesto que además de influir en el control del ciclo celular, también se involucra en la integridad del ADN así como en la supervivencia de las células expuestas a agentes que dañan el mismo.

A pesar de sus múltiples funciones beneficiosas para el cuerpo humano, su respectiva alteración transfiere un riesgo muy elevado al desarrollo del cáncer. Bajo condiciones normales, cuando se produce un daño en el ADN, el gen TP53 detiene el ciclo celular, no obstante, en caso de que la proteína p53 esté mutada, el ciclo celular continúa y el ADN dañado se replica; en consecuencia, se produce una proliferación celular descontrolada[1], lo cual desemboca en tumores cancerosos. Respecto a lo anterior, Patiño et al. (2004, p. 88) agregan que este gen "se encuentra alterado en cerca del 60% de todos los tipos de tumores", asimismo, estiman que "aproximadamente de los 6.5 millones de casos de cáncer informados anualmente en el mundo, 2.4 millones de los mismos ocurren por mutaciones de dicho gen".

En conocencia de lo expuesto en los párrafos anteriores, se propone el estudio de la expresión del gen versado, en el tejido mamario de los pacientes y de las personas en riesgo de cáncer de pecho, el cual "es el tipo de cáncer más frecuente y la causa más común de muerte por cáncer en mujeres a nivel mundial" (Organización Panamericana de la Salud, s.f.). Asimismo, se formula un análisis segmentado por grupos etarios: la adolescencia, que comprende de los 12 a los 18 años; la juventud, que agrupa a los individuos de entre 19 y 30 años; la adultez, que va de los 30 a los 60 años y finalmente la vejez, de 60 años en adelante.

2 Delimitación del tema

El proyecto se delimita en torno a dos ejes temáticos que se fusionan: el gen TP53 y el cáncer de pecho. De esta forma, se vincula la presencia del primero con el desarrollo del segundo. Asimismo, el trabajo se segmentará en los grupos etarios definidos previamente.

Por otro lado, en un primer momento se analizarán las expresiones del gen TP53 en el tejido mamario de pacientes y de personas en riesgo de cáncer de pecho, a partir de la base internacional seleccionada. Una vez contruidas las conclusiones pertinentes, las mismas se trasladarán al escenario costarricense mediante el acceso a la base de datos del Ministerio de Salud relacionada al recuento de personas portadoras de cáncer de pecho. Lo anterior con el objetivo de formular ciertas hipótesis sobre el vínculo existente entre la expresión del gen mencionado y el desarrollo del cáncer de pecho en Costa Rica.

[1] La proliferación celular descontrolada consiste en el crecimiento y la división de las células a una velocidad anormal, sin responder a las señales de detención o muerte celular.

3 Bases de datos

La principal base de datos a utilizarse corresponde a la proporcionada por el Instituto Nacional de Cáncer en Estados Unidos: “The TP53 Dataset, R21”. La versión R21 es la más reciente, se actualizó en enero de 2025. Este recurso contiene alrededor de 29,900 variantes de tumores, más de 2,155 individuos con variantes del gen *TP53* confirmadas y datos funcionales sobre más de 9,000 proteínas mutantes. En particular, la información se filtrará y se tomará únicamente aquella en donde la morfología coincida con los pechos, es decir, que las anormalidades se presenten en el área estipulada.

En el caso de Costa Rica, la base de datos del Ministerio de Salud se utilizará en una vía mayormente cualitativa, orientada a la formulación de hipótesis significativas en la materia. Dicha herramienta se titula: “Incidencia de tumores malignos diferentes características” y su versión más reciente corresponde a la del año 2022; asimismo, segmenta el recuento por sexo y rangos de edad.

Primeramente, se va a realizar un filtro para obtener las columnas y filas de interés de la base de datos internacional, como se muestra a continuación. Se trabajará con la misma a partir de aquí.

```
library(readr)

library(dplyr)
```

```
Adjuntando el paquete: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(ggplot2)

url <- "https://raw.githubusercontent.com/eetefy2311/TP53-data/refs/heads/main/GermlineDownloa

temp.file <- tempfile(fileext = ".csv")

download.file(url, destfile = temp.file, mode = "wb")

base.de.datos <- read_csv(temp.file)
```

```
Rows: 4649 Columns: 74

-- Column specification --------------------------------------------------------
Delimiter: ","
```

```
chr (45): Family_code, Country, Population, Region, Development, Class, Germ...
dbl (27): Family_ID, Generations_analyzed, MUT_ID, hg18_Chr17_coordinates, h...
lgl  (2): Dead, Unaffected

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
View(base.de.datos)
```

3.1 Caraterísticas de los datos

En primera instancia, es primordial resaltar que la tabla de datos encargada de solventar la mayor parte del proyecto es aquella que se genera a partir de la base de datos internacional, ya que con ella se busca determinar tendencias en las expresiones del gen TP53 presentes en el tejido mamario de pacientes o personas que potencialmente podrían desarrollar cáncer de pecho. En contraste, tal como se mencionó antes, los datos extendidos por el Ministerio de Salud de Costa Rica se emplearán desde un enfoque cualitativo, de forma que estos contribuyan en la formulación de hipótesis relacionadas a dicho tema para la población nacional. En motivo de lo anterior, el presente apartado además de los sucesores, se apoyarán solo en la tabla de información filtrada de la base de datos internacional.

De esta forma, se emplearon dos paramétros clave para filtrar la tabla de datos: el primero se basó en la utilidad de las variables de estudio para el proyecto y el segundo consistió en que la topografía fuera igual en todos los casos al área del pecho, es decir, que los padecimientos se situaran en el sector mencionado. A partir de ello, se generó una tabla de datos que contiene la información de 1,153 pacientes en su mayoría del sexo femenino, de nacionalidades variadas y con edades que oscilan entre los 14 y los 90 años.

La totalidad de la información fue recolectada por el Instituto Nacional de Cáncer de Estados Unidos y la última actualización de la base de datos fue en enero de 2025. Tales actualizaciones, especialmente las que conciernen a las variantes del gen TP53, se sustentan de los datos reportados en la literatura médica publicada, al igual que de otras bases de datos públicas.

3.2 Población de estudio

La población de estudio incluye pacientes y personas con un riesgo elevado de desarrollar cáncer de pecho, con edades comprendidas entre los 14 y los 90 años, que además presentan expresiones del gen TP53 en el tejido mamario. Los datos de los individuos fueron recolectados por el Instituto Nacional de Cáncer de Estados Unidos en enero de 2025.

3.3 Muestra observada

La muestra que se seleccionó está conformada por un total de 1,153 individuos de edades y nacionalidades variadas, cuyos datos fueron recolectados por el Instituto Nacional de Cáncer de Estados Unidos en enero de 2025. Tales sujetos poseen expresiones del gen TP53 en el tejido mamario, al mismo tiempo que son pacientes de cáncer de pecho o poseen un alto riesgo de contraerlo.

3.4 Unidad estadística o individuos

La unidad estadística se define en cada uno de los individuos que padecen o poseen un alto riesgo de desarrollar cáncer de pecho, que muestran expresiones del gen TP53 en el tejido mamario y cuyos datos fueron recolectados por el Instituto Nacional de Cáncer de Estados Unidos en enero de 2025.

3.5 Variables de estudio

- **País** (Country, cualitativa): país del que proviene la persona.
- **Población** (Population, cualitativa): población de la que proviene la persona.
- **Región** (Region, cualitativa): región de la que proviene la persona.
- **Gen mutado** (Germline_mutation, cualitativa): tiene un valor único, el gen TP53.
- **Identificador personal** (Individual_ID, cualitativa): identificador individual de cada persona.
- **Sexo** (Sex, cualitativa): corresponde al sexo de la persona. F se denota para femenino y M para masculino.
- **Edad de diagnóstico** (Age_at_diagnosis, cuantitativa): edad en la que fue diagnosticado el padecimiento en la persona.
- **Topografía** (Topography, cualitativa): área en la que se encuentra el padecimiento.
- **Topografía específica** (Short_topo, cualitativa): área específica en la que se encuentra el padecimiento.
- **Morfología** (Morphology, cualitativa): se refiere al tipo de padecimiento.
- **Muerte** (Dead, cualitativa): indica si la persona está muerta o viva. Se denota TRUE cuando está muerta y FALSE cuando está viva.

4 Primeras filas de la tabla de datos

Para esta sección, se utiliza head() para observar las primeras 5 filas de la base de datos.

```
base.de.datos.filtrada <- base.de.datos %>%
  select(Country, Population, Region, Germline_mutation,
         Individual_ID, MUT_ID, TCGA_ICGC_GENIE_count,
         Sex,Age_at_diagnosis, Topography, Short_topo, Morphology,
         Dead, Age) %>%
  filter(Topography == "BREAST")

resumen.datos <- base.de.datos.filtrada %>%
  head(5)
resumen.datos
```

```
# A tibble: 5 x 14
  Country Population      Region Germline_mutation Individual_ID MUT_ID
  <chr>   <chr>           <chr>  <chr>                     <dbl>  <dbl>
1 UK      Northern Europe Europe TP53                          6   2821
2 UK      Northern Europe Europe TP53                         25   3297
3 UK      Northern Europe Europe TP53                         24   3297
4 UK      Northern Europe Europe TP53                         24   3297
5 UK      Northern Europe Europe TP53                        570   2143
# i 8 more variables: TCGA_ICGC_GENIE_count <dbl>, Sex <chr>,
#   Age_at_diagnosis <dbl>, Topography <chr>, Short_topo <chr>,
#   Morphology <chr>, Dead <lgl>, Age <dbl>
```

5 Análisis del resumen de las variables cuantitativas

```r
library(dplyr)

#Se genera un resumen de la única variable cuantitativa usando el paquete base.
summary(base.de.datos.filtrada$Age_at_diagnosis)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  14.00   28.00   34.00   36.24   42.00   90.00     178
```

Análisis del resumen

En primera instancia, como se ha mencionado en múltiples ocasiones, la edad de diagnóstico oscila entre los 14 y lo 90 años, lo que ofrece un amplio panorama en términos etarios para desarrollar el análisis segmentdo. Después, la mediana es de 34 años: la mitad de los casos de la tabla filtrada se dieron antes de esta edad; un indicador de que la mayoría de los padecimientos diagnosticados se presentan en personas jóvenes. Por su parte, el promedio es ligeramente mayor, de tal forma que se sugiere la existencia de edades considerablemente altas que desplazan la media hacia la derecha. Finalmente, en lo que respecta a los cuartiles, se tiene que un 25% de los padecimientos se diagnosticó antes de los 28 años y un 75% se diagnosticó antes de los 42 años, reafirmando el hecho de que la mayoría de los casos de la tabla de datos se presenta en personas de una edad no tan avanzada.

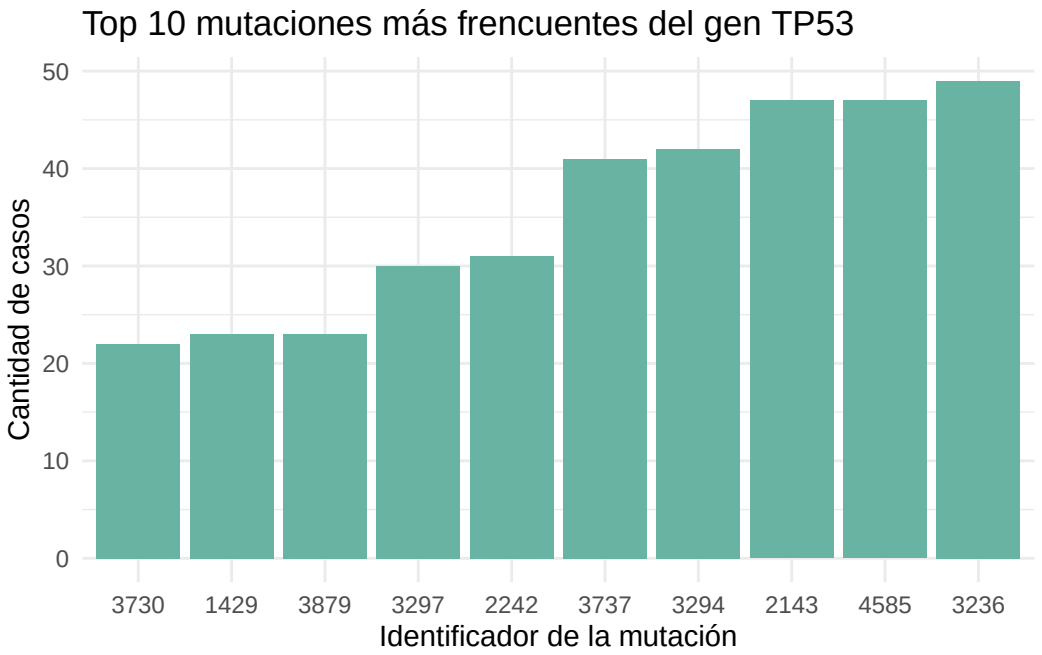
6 Gráficos que describe la distribución de las variables cuantitativas

Las variables cuantitativas que se estarán tomando en cuenta son: -“MUT_ID”: se refiere al tipo de mutación del gen identificado con un número único. Con el mismo, se plantea ver diferenctes frecuencias de mutaciones en relación con otras variables. -“Age_at_diagnosis”: se refiere al año en que los pacientes fueron diagnosticados con cáncer, de disntios tipos (de acuerdo a su topología). -“Age”: se refiere a la edad de los pacientes. En general, se nota que hay un grado alto de “NA” en esta columna. Note que a la par de esta hay una columna que indica si el paciente esta vivo o muerto, dado que en gran medida los paicentes con cáncer no logran sobrevivir, es por esto que la columna de edades tiene en su mayoría valores nulos. -“TCGA_ICGC_GENIE_count”: se refiere a la cuenta de casos de cáncer medidos por estas entidades en

conjunto, donde TCGA significa “The Cancer Genome Atlas”, ICGC significa “International Cancer Genome Consortium” y GENIE significa “Genomics Evidence Neoplasia Information Exchange”. Todas son organizaciones que se dedican a la difusión de bases de datos refrentes a pacientes con cáncer.

Para la elaboración de los graficos, dado que todas las variables son numéricas, se utiliza un for para generar ls gráficos por cada variable que se identificó previamente.

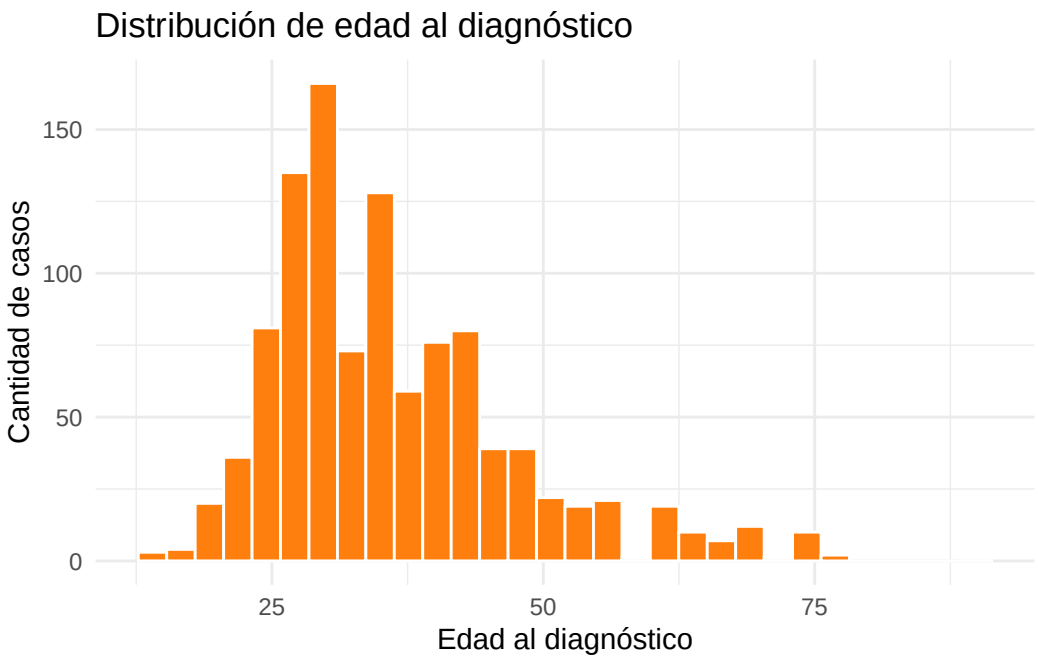
```
vararіables.cuantitativas <- c("MUT_ID","Age_at_diagnosis", "Age", "TCGA_ICGC_GENIE_count")

#Grafico de las 10 mutaciones más frecuentes en la base de datos

top10 <- base.de.datos.filtrada %>%
  group_by(MUT_ID) %>%
  summarize(n = n()) %>%
  arrange(desc(n)) %>%
  head(10)

ggplot(top10, aes(x = reorder(MUT_ID, n), y = n)) +
  geom_col(fill = "#69b3a2") +
  labs(title = "Top 10 mutaciones más frencuentes del gen TP53",
       x = "Identificador de la mutación",
       y = "Cantidad de casos") +
  theme_minimal()
```




```
#Grafico del año de diagnóstico
ggplot(base.de.datos.filtrada, aes(x = Age_at_diagnosis)) +
  geom_histogram(bins = 30, fill = "#ff7f0e", color = "white") +
  labs(title = "Distribución de edad al diagnóstico",
```

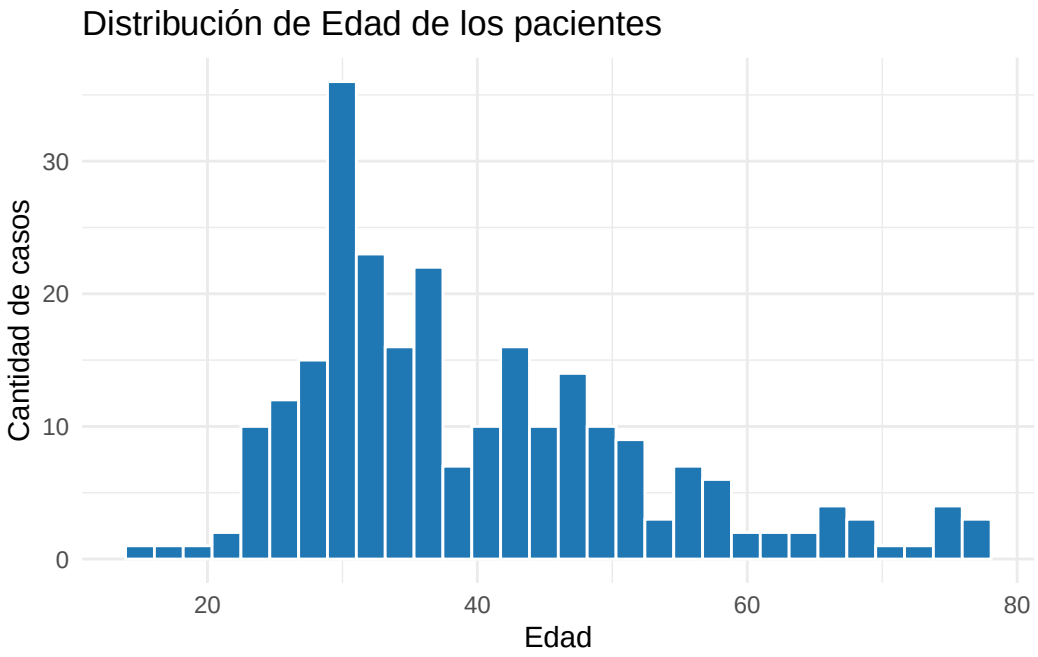
```
       x = "Edad al diagnóstico",
       y = "Cantidad de casos") +
  theme_minimal()
```

```
Warning: Removed 178 rows containing non-finite outside the scale range
(`stat_bin()`).
```

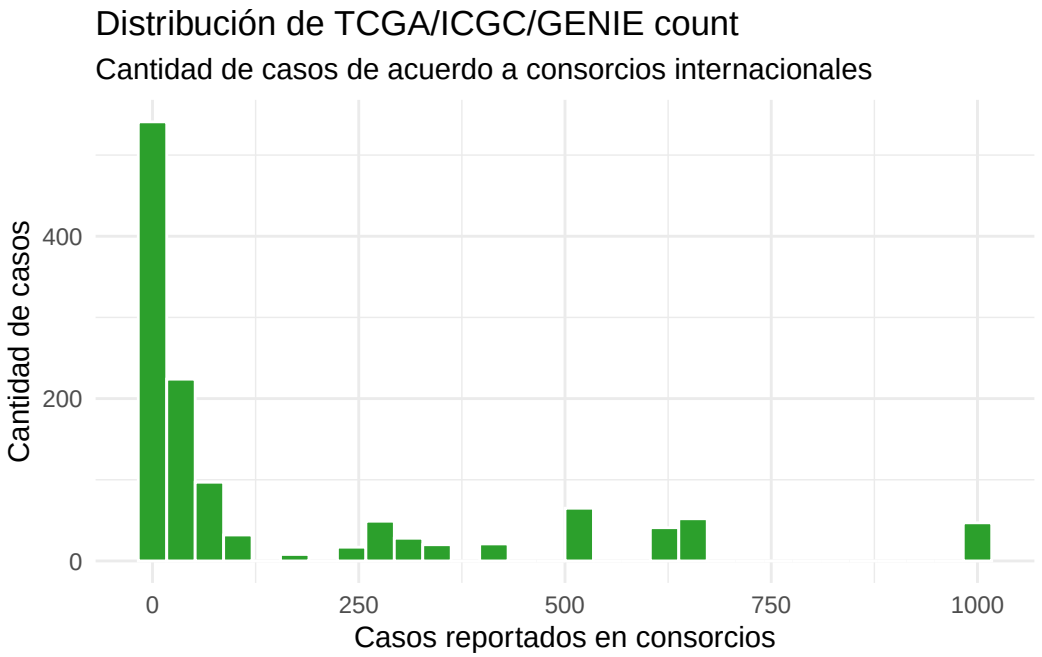



```
#Gráfico de edad actual de los pacientes

ggplot(base.de.datos.filtrada, aes(x = Age)) +
  geom_histogram(bins = 30, fill = "#1f77b4", color = "white") +
  labs(title = "Distribución de Edad de los pacientes",
       x = "Edad",
       y = "Cantidad de casos") +
  theme_minimal()
```

```
Warning: Removed 989 rows containing non-finite outside the scale range
(`stat_bin()`).
```




```
#Gráfico de la cuenta de casos de acuerdo a organizaciones
#internacionales anteriormente mencionadas

ggplot(base.de.datos.filtrada, aes(x = TCGA_ICGC_GENIE_count)) +
  geom_histogram(bins = 30, fill = "#2ca02c", color = "white") +
  labs(title = "Distribución de TCGA/ICGC/GENIE count",
       subtitle = "Cantidad de casos de acuerdo a consorcios internacionales",
       x = "Casos reportados en consorcios",
       y = "Cantidad de casos") +
  theme_minimal()
```



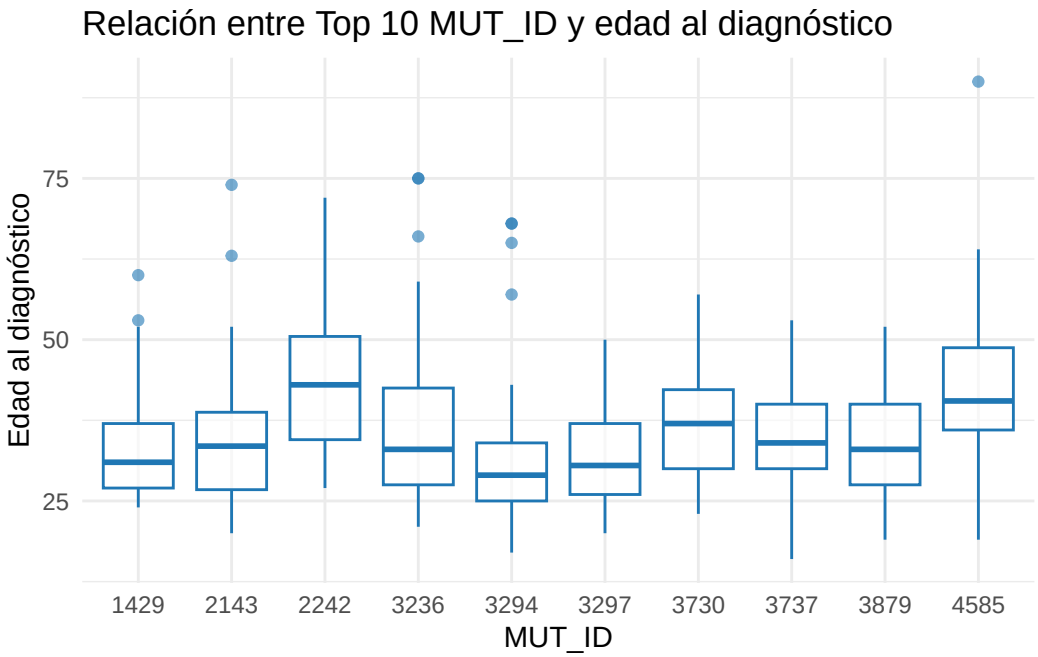
7 Gráficos que describen la relación entre las variables

Para esta parte, se plantea la descripción de dos relaciones: -Año de diagnóstico y tipo de mutación: indica si hay alguna mutación que sea más frencuente en un grupo estario específico. -Casos en TCGA/ICGC/GENIE y la edad de diagnóstico: se plantea determinar si existe una relación entre la cantidad de casos y la edad.

De nuevo, se han a utilizar los primeros 10 valores más frecuentes de cada variable en la base de datos, para observar mejor su relación

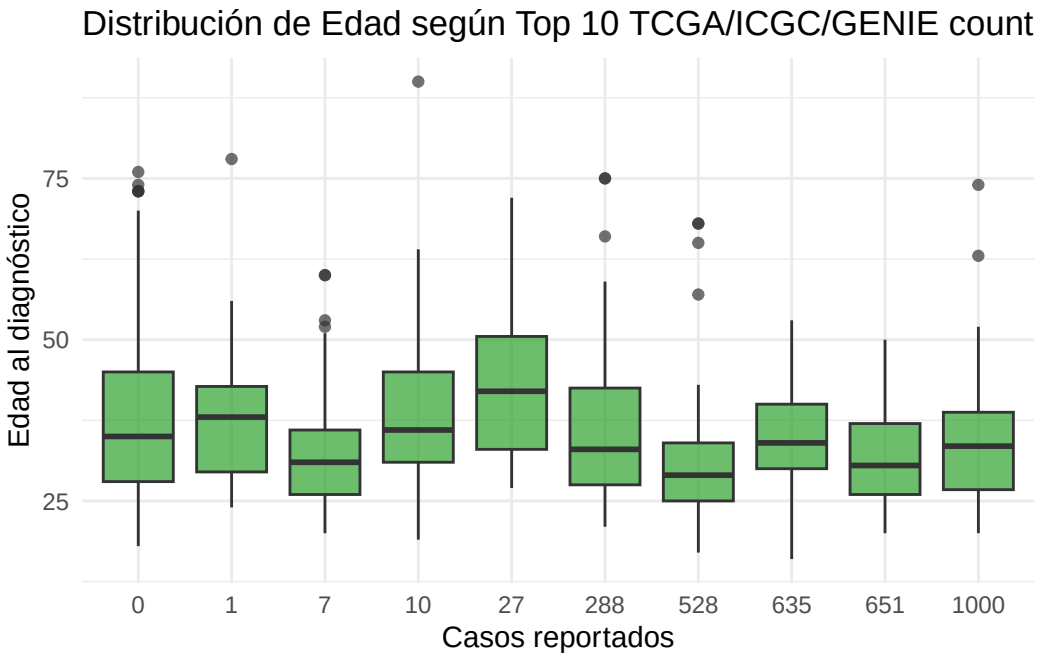
```
top.mut <- base.de.datos.filtrada %>%
  group_by(MUT_ID) %>%
  summarize(n = n()) %>%
  arrange(desc(n)) %>%
  head(10)

base.de.datos.filtrada %>%
  filter(MUT_ID %in% top.mut$MUT_ID) %>%
  ggplot(aes(x = as.factor(MUT_ID), y = Age_at_diagnosis)) +
  geom_boxplot(alpha = 0.6, color = "#1f77b4") +
  labs(title = "Relación entre Top 10 MUT_ID y edad al diagnóstico",
       x = "MUT_ID",
       y = "Edad al diagnóstico") +
  theme_minimal()
```

```
Warning: Removed 41 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```




```
top.casos <- base.de.datos.filtrada %>%
  group_by(TCGA_ICGC_GENIE_count) %>%
  summarize(n = n()) %>%
  arrange(desc(n)) %>%
  head(10)

base.de.datos.filtrada %>%
  filter(TCGA_ICGC_GENIE_count %in% top.casos$TCGA_ICGC_GENIE_count) %>%
  ggplot(aes(x = as.factor(TCGA_ICGC_GENIE_count), y = Age_at_diagnosis)) +
  geom_boxplot(fill = "#2ca02c", alpha = 0.7) +
  labs(title = "Distribución de Edad según Top 10 TCGA/ICGC/GENIE count",
       x = "Casos reportados",
       y = "Edad al diagnóstico") +
  theme_minimal()
```

```
Warning: Removed 77 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```

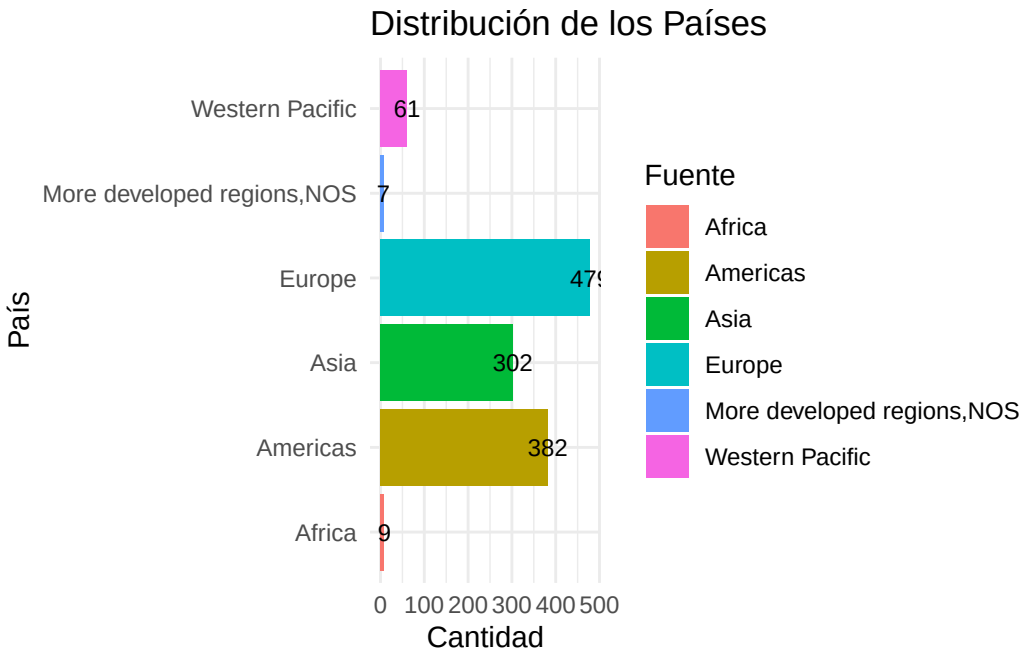


8 Gráfico de la distribución de las variables categóricas

Para mostrar la distribución de una de las variables categórcas, se seleccionó la variable Región/País, como se muestra a continuación.

```
base.de.datos.filtrada %>%
  filter(is.na(Region) == FALSE) %>%
  ggplot(aes(x = Region, fill = Region)) +
  geom_bar() +
  coord_flip() +
  guides(fill = guide_legend(title = "Fuente")) +
  labs(title = "Distribución de los Países",
       x = "País",
       y = "Cantidad") +
  geom_text(aes(label = ..count..), stat = "count", vjust = 0.5, colour = "black", size = 3) +
  theme_minimal()
```

```
Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
i Please use `after_stat(count)` instead.
```



9 Identificación de los valores faltantes y los posibles outliers

Para la identificar valores faltantes y outliers, se utiliza la regla 1,5 x IQR. De manera que, se encuentra el procentaje de datos nulos que hay por columna y la cantidad de outliers de los valores numericos.

```
valores.nulos <- colSums(is.na(base.de.datos.filtrada))
tamanio <- nrow(base.de.datos.filtrada)
porcentaje.por.columna <- valores.nulos / tamanio * 100
porcentaje.por.columna
```

```
             Country           Population               Region
           0.1610306            0.1610306            0.1610306
    Germline_mutation         Individual_ID               MUT_ID
           0.0000000            0.0000000            0.0000000
TCGA_ICGC_GENIE_count                  Sex      Age_at_diagnosis
           0.0000000            0.0805153           14.3317230
           Topography           Short_topo           Morphology
           0.0000000            0.0000000            0.0000000
                Dead                  Age
           0.0000000           79.6296296
```

```
total.outliers <- 0

# Usando la regla de 1*5 x IQR (identificador de outliers)

for (i in 1:ncol(base.de.datos.filtrada)) {
```

```
  x <- base.de.datos.filtrada[[i]]

  if (!is.numeric(x)) next

  q <- quantile(x, probs = c(0.25, 0.5, 0.75), na.rm = TRUE)
  q1 <- q[1]
  q2 <- q[2]
  q3 <- q[3]

  iqr <- q3 - q1
  limite_inferior <- q1 - 1.5 * iqr
  limite_superior <- q3 + 1.5 * iqr

  outliers <- x[x < limite_inferior | x > limite_superior]
  total.outliers <- total.outliers + length(outliers)
}

total.outliers
```

```
[1] 1324
```

10 Técnicas para subsanar los valores perdidos y los outliers

Según Ferro (2024), dentro de las maneras disponibles para tratar con valores faltantes, está omitir los datos con la función na.omit() la cual devuelve un nuevo dataframe eliminando las filas con valores nulos. Este método se utiliza cuando podría tratarse de un error, o los datos involucrados no son muy relevantes para el objetivo del análisis que se busca aplicar.

En caso de ser necesario contar con esos datos para dar una predicción, se puede imputar los datos escogiendo cambiar por una media para mayor simplicidad, o por un valor experto proveniente de una persona que conoce acerca del área y tiene experiencia en la observación de la manifestación de estos casos. También, se puede deducir mediante alguna relación de fórmula con las demás categorías.

Según VSNi (2021), una manera de tratar con los valores atípicos es transformar los datos de acuerdo al tipo de distribución al cual tiendan. En el caso de que su distribución se acerca más a número altos por la derecha, utilizar una transformación logarítmica que incluya los valores altos que se manifiestan. En el caso de que el valor atípico pueda ser un error, es mejor eliminarlo. Adicionalmente, se puede estudiar el modelo de distribución para encontrar la probabilidad de que se haya manifestado ese valor atípico y aplicar en distintos casos.

Referencias

Cave, V. (2024). *Handling outliers in statistical analysis: To remove or not to remove?* VSNi. https://vsni.co.uk/should-i-drop-outliers/

Ferro Alfonso, L. A. (2024). *Limpieza, organización y preparación de datos en R*. RPubs. https://rpubs.com/profe_ferro/1213826

Herrera Patiño, J., Vásquez Palacio, G., Ramírez Castro, J. y Muñetón Peña, C. (2004). Papel del gen TP53 en la oncogénesis. *Salud UIS, 36*, 88-99. https://dialnet.unirioja.es/servlet/articulo?codigo=8217251

Instituto Nacional de Cáncer de Estados Unidos. (2025). *The TP53 Database*. Instituto Nacional de Cáncer de Estados Unidos. https://tp53.cancer.gov/

López M., Anzola, M., Cuevas-Salazar, N., Aguirre, J. y Martínez, M. (2001). p53, un gen supresor tumoral. *Gaceta Médica de Bilbao*, 98(1), 21–27. https://www.gacetamedicabilbao.eus/index.php/gacetamedicabilbao/article/view/596

Organización Panamericana de la Salud. (s.f.). *Cáncer de mama*. Organización Panamericana de la Salud. https://www.paho.org/es/temas/cancer-mama