```
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

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.
```

1 Fundamentación del proyecto

1.1 Pregunta de investigación

¿Cómo varía la expresión y la frecuencia de las mutaciones del gen TP53 en el tejido mamario de pacientes y personas en riesgo de desarrollar cáncer de pecho, considerando los distintos grupos etarios? ¿Cómo se relacionan los resultados obtenidos a los registrados en Costa Rica?

2 Objetivos

2.1 Objetivo general

Analizar la expresión y frecuencia de las mutaciones del gen TP53 en el tejido mamario de pacientes y personas en riesgo de desarrollar cáncer de pecho, diferenciando los resultados por grupos etarios.

2.2 Objetivos específicos

- Identificar las mutaciones más frecuentes del gen TP53 presentes en el tejido mamario de los individuos analizados.
- Describir la distribución etaria de los casos diagnosticados con mutaciones en el gen TP53.
- Determinar la relación existente entre los tipos de mutación y la edad al diagnóstico.
- Comparar la prevalencia de casos según las regiones geográficas y consorcios internacionales de origen.
- Formular hipótesis preliminares sobre la posible correlación entre las alteraciones del gen TP53 y la incidencia de cáncer de pecho en Costa Rica, a partir de los registros del Ministerio de Salud.

3 Hipótesis

3.1 Hipótesis general:

La expresión del gen TP53 en el tejido mamario presenta variaciones significativas según el grupo etario, siendo más frecuentes las mutaciones en adultos jóvenes y adultos medios.

4 Hipótesis específicas:

- H1: Las mutaciones más frecuentes del gen TP53 se concentran en individuos diagnosticados entre los 30 y 50 años.
- H2: Existe una relación significativa entre el tipo de mutación (MUT_ID) y la edad al diagnóstico (Age_at_diagnosis).
- H3: Las regiones con mayor representación de casos presentan también mayor diversidad mutacional del gen TP53.

(Agregar fuentes)

5 Justificación

El estudio del gen TP53 es de vital importancia para la comprensión de los mecanismos moleculares del cáncer de mama. Las mutaciones en este gen pueden conducir a una proliferación celular descontrolada y a la pérdida de la función supresora tumoral, incrementando el riesgo de malignidad. Analizar su expresión por grupos etarios permitirá identificar patrones que podrían asociarse a factores de riesgo biológicos o ambientales específicos.

Además, la comparación con datos del Ministerio de Salud de Costa Rica ofrece la posibilidad de contextualizar los hallazgos en la realidad nacional, contribuyendo al diseño de estrategias de prevención y detección temprana más precisas.

6 Marco metodológico

El presente estudio se desarrollará bajo un enfoque cuantitativo de tipo descriptivo. Su propósito será analizar cómo varía la expresión y la frecuencia de las mutaciones del gen TP53 en el tejido mamario de pacientes y personas en riesgo de desarrollar cáncer de pecho, tomando en cuenta los diferentes grupos etarios. Este diseño permitirá observar las relaciones existentes entre variables genéticas, demográficas y clínicas sin manipularlas directamente, centrándose en la descripción, comparación y correlación de los datos obtenidos.

La investigación se basará en el uso de bases de datos secundarias de carácter público y científico. La principal fuente será la base internacional The TP53 Dataset (versión R21), desarrollada por el Instituto Nacional de Cáncer de los Estados Unidos, la cual contiene información sobre más de 29,000 variantes tumorales y más de 2,000 individuos con mutaciones confirmadas en el gen TP53. De esta base se seleccionarán los registros cuya topografía corresponda al tejido mamario, a fin de centrar el análisis en los casos de cáncer de mama. Complementariamente, se considerarán los registros nacionales provenientes del Ministerio de Salud de Costa Rica (2022), con el objetivo de contrastar los hallazgos internacionales con la incidencia local del cáncer de mama.

En una primera etapa, se aplicarán estadísticas descriptivas con el fin de caracterizar la población en estudio. Se calcularán medidas de tendencia central (media, mediana y moda) y de dispersión (desviación estándar y rango intercuartílico) para variables como la edad al diagnóstico y la frecuencia de casos reportados. Asimismo, se elaborarán tablas de frecuencia, histogramas y diagramas de caja, que permitirán visualizar la distribución de las edades, la variación de las mutaciones y la presencia de posibles valores atípicos. Este análisis inicial ayudará a comprender la composición general de la muestra y a detectar patrones preliminares en la distribución de las mutaciones según la edad.

Posteriormente, se llevará a cabo un análisis comparativo entre grupos, con el propósito de determinar si existen diferencias significativas en la edad al diagnóstico según el tipo de mutación del gen TP53. Para ello se aplicará una prueba de análisis de varianza (ANOVA), en caso de que los datos cumplan con los supuestos de normalidad y homogeneidad de varianzas. Si dichos supuestos no se cumplen, se recurrirá a la prueba no paramétrica de Kruskal–Wallis, la cual permitirá contrastar las medianas de edad entre grupos sin asumir una distribución normal. Este análisis facilitará identificar si ciertos tipos de mutación se asocian a edades de diagnóstico más tempranas o tardías.

(Añadir referencias)

Además, se aplicarán análisis de correlación con el fin de evaluar la relación existente entre la edad al diagnóstico (Age_at_diagnosis) y el número de casos reportados por los principales consorcios internacionales (TCGA, ICGC y GENIE). Dependiendo de la distribución de los datos, se utilizarán los coeficientes de correlación de *Pearson* (para datos normales) o *Spearman* (para datos no normales). Estos análisis permitirán establecer si la edad al diagnóstico se relaciona de manera significativa con la frecuencia o severidad de las mutaciones del gen TP53, aportando evidencia empírica a la hipótesis de que la edad influye en la expresión de dicho gen.

Asimismo, se llevará a cabo un análisis de frecuencias y proporciones para determinar la incidencia relativa de las mutaciones por grupo etario y por tipo de mutación. Esta técnica permitirá identificar cuáles grupos de edad presentan mayor prevalencia de mutaciones en el gen TP53, lo que resultará clave para establecer patrones de riesgo y posibles implicaciones clínicas.

Antes de aplicar los procedimientos inferenciales, se realizará una limpieza y depuración de los datos (Esta se trabajó previamente en la Bitácora 2). Este proceso incluirá la identificación de valores faltantes, que serán tratados mediante eliminación o imputación según su relevancia; la detección de valores atípicos a través del rango intercuartílico (IQR), para garantizar la coherencia interna del conjunto de datos. Este paso será fundamental para asegurar la calidad y fiabilidad de los análisis posteriores.

De este modo, cada método estadístico aportará evidencia específica para responder la pregunta de investigación:

- Los análisis descriptivos permitirán conocer cómo se distribuyen la edad y las mutaciones del gen TP53.
- Las pruebas de comparación (ANOVA/Kruskal–Wallis) identificarán si existen diferencias significativas en la edad de diagnóstico según el tipo de mutación.
- Los análisis de correlación mostrarán si existe una relación entre la edad y la frecuencia de mutaciones reportadas.
- Finalmente, el análisis de frecuencias permitirá observar la incidencia relativa de las mutaciones en cada grupo etario.

Posteriormente, se llevará a cabo un análisis comparativo entre regiones geográficas, tomando como

base la variable Country de la base internacional. Se describirá la distribución de mutaciones del gen TP53 en diferentes países y regiones, con especial énfasis en América Latina. Esta información se contrastará cualitativamente con la base de datos del Ministerio de Salud de Costa Rica, para explorar si los patrones internacionales guardan relación con las tasas de incidencia y mortalidad por cáncer de mama en la población costarricense. Aunque no se establecerán correlaciones estadísticas directas entre ambas bases debido a la diferencia en el tipo de información, este ejercicio permitirá contextualizar los resultados globales en el ámbito nacional, y formular hipótesis sobre la posible prevalencia de mutaciones similares en el país.

Frecuencias de MUT_ID

```
frecuencias.mutaciones <- base.de.datos %>%
  group_by(MUT_ID) %>%
  summarise(
    frecuencia = n(),
    porcentaje = round((n()/nrow(base.de.datos))*100, 3)
  ) %>%
  arrange(desc(frecuencia))
```

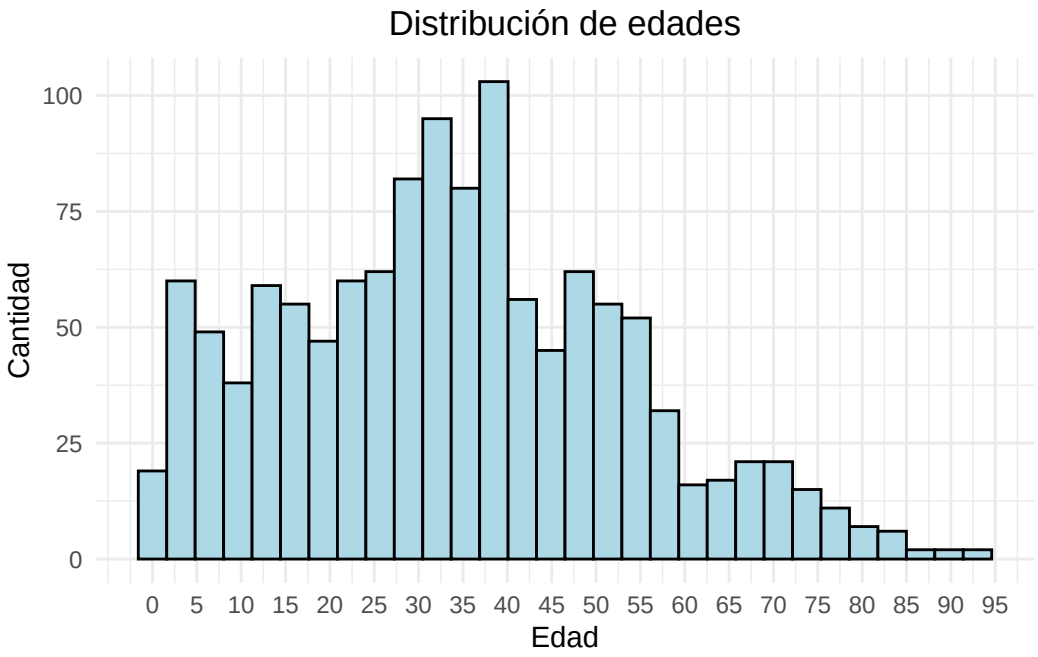
Top 10 Mutaciones con mayor frecuencia

```
top.mutaciones <- frecuencias.mutaciones %>%
  head(10)
```

Histograma sobre la distribución de edades

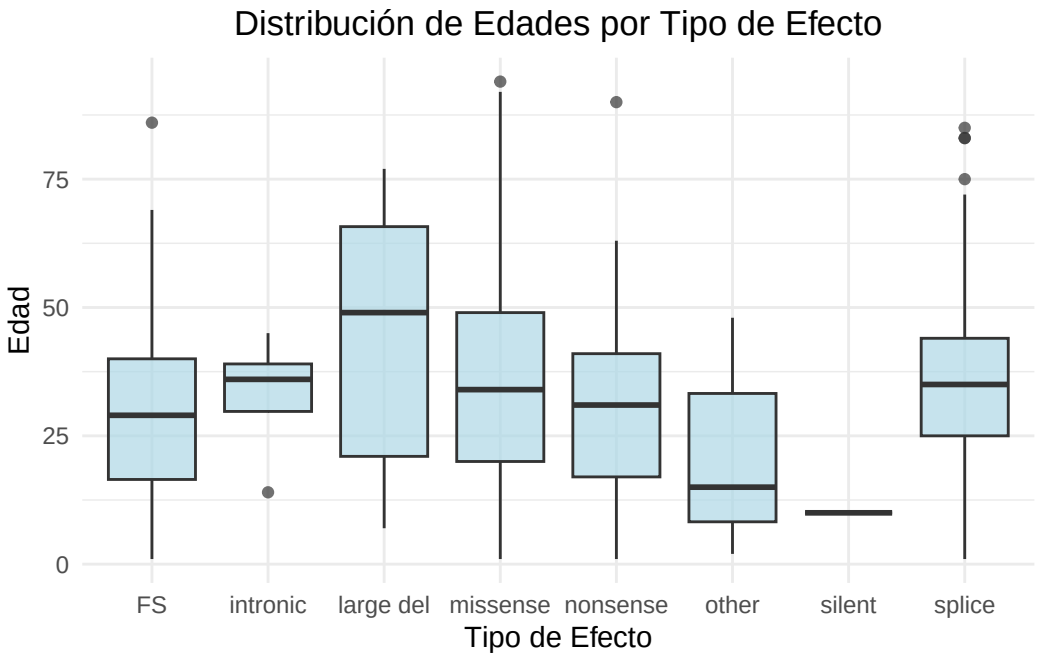
```
base.de.datos %>%
  filter(!is.na(Age)) %>%
  ggplot(aes(x = Age)) +
  geom_histogram(color = "black", fill = "lightblue") +
  scale_x_continuous(breaks = seq(0,100,5)) +
  labs(title = "Distribución de edades",
       x = "Edad",
       y = "Cantidad") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



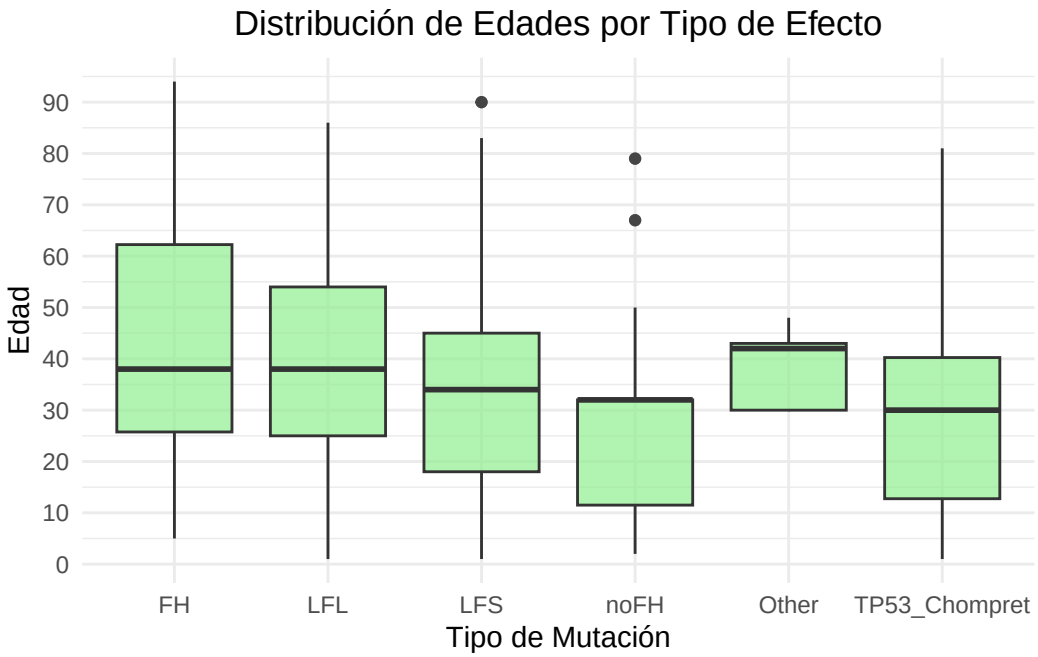
Boxplot de Edades con respecto al Efecto de Mutación

```
base.de.datos %>%
  filter(!is.na(Age), !is.na(Effect)) %>%
  ggplot(aes(x = Effect, y = Age)) +
  geom_boxplot(fill = "lightblue", alpha = 0.7) +
  labs(title = "Distribución de Edades por Tipo de Efecto",
       x = "Tipo de Efecto",
       y = "Edad") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



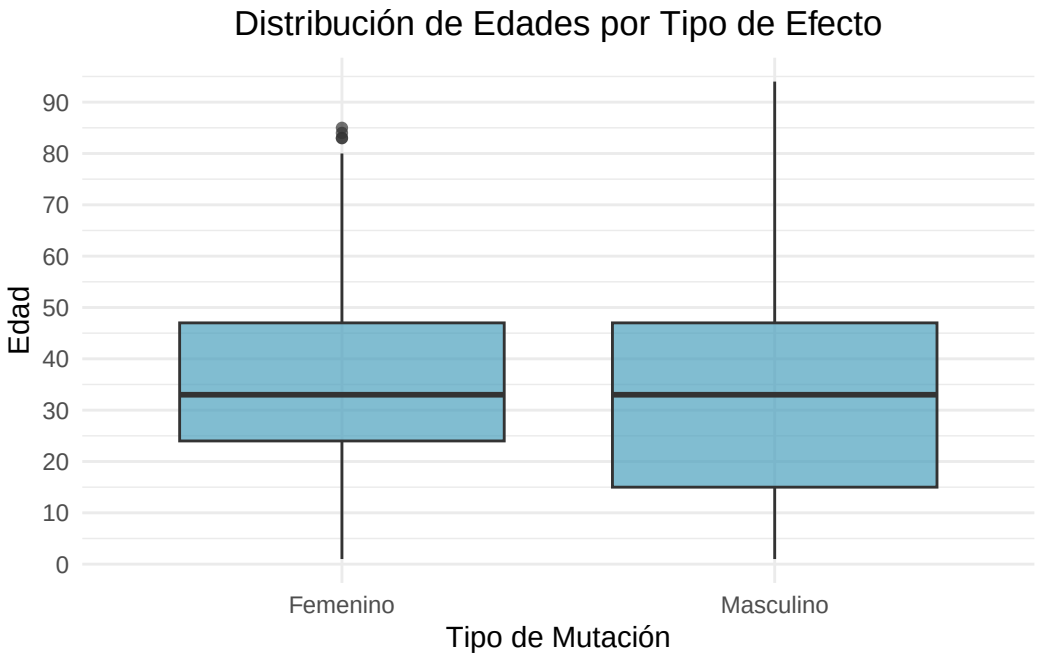
Boxplot de Edades con respecto a la Clase de Mutación

```
base.de.datos %>%
  filter(!is.na(Age), !is.na(Class)) %>%
  ggplot(aes(x = Class, y = Age)) +
  geom_boxplot(fill = "lightgreen", alpha = 0.7) +
  scale_y_continuous(breaks = seq(0,100,10)) +
  labs(title = "Distribución de Edades por Tipo de Efecto",
       x = "Tipo de Mutación",
       y = "Edad") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



Boxplot de Edades y Sexo

```
base.de.datos %>%
  filter(!is.na(Age), !is.na(Sex)) %>%
  mutate(Sexo = ifelse(Sex == "F", "Femenino", "Masculino")) %>%
  ggplot(aes(x = Sexo, y = Age)) +
  geom_boxplot(fill = "#4AA0BD", alpha = 0.7) +
  scale_y_continuous(breaks = seq(0,100,10)) +
  labs(title = "Distribución de Edades por Tipo de Efecto",
       x = "Tipo de Mutación",
       y = "Edad") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



Estadísticas Descriptivas de la Edad

Edad promedio: 34 Edad mediana: 33 Edad máxima: 94 Edad mínima: 1

Primer cuartil: 19 Segundo cuartil: 33 Tercer cuartil: 47

Desviación estándar: 19.2 Varianza: 369 IQR: 28

Conclusiones

Nuestra primera hipósis hipótesis fue que las edades estaban concentradas en un rango entre los 30 y 50 años, basados en los resultados obtenidos en el gráfico boxplot, se puede llegar a la conclusión de que el rango de edad concentrado entre los 55 y 20 años, con mayores frecuencias entre los 30 y 40 años. Por otro lado se obtuvo que la edad promedio fue 34, mientras que la mediana de 33, lo que indica que no hay un sesgo a edades muy jóvenes o muy avanzadas; también se puede decir que en general las mutaciones del gen TP53 no se está en gran medida en pacientes de más de 50 años, sino que se está concentrando en edades de menos de 50.

Además, se obtuvo que la mutación más frecuente fue la del identificador 4585 con 314 observaciones, representando un 6.754% del total de observaciones, casi el doble que la mutación en el segundo lugar, la cual tiene el identificador 2143 con 180 observaciones, que representa un 3.872%, y la tercer mutación con mayor frecuencia fue el 3236 con 160 observaciones, representando un 3.442% del total de mutaciones.