MEMORIA.

En primera instancia más o menos sabemos que vamos a encontrar bastantes variables categóricas en el dataframe con los datos clínicos (el género del paciente, si ha recibido radioterapia, y por el contrario, el de los datos de secuenciación sólo tendrá valores numéricos (de la expresión de los genes).

Cabe destacar que los valores de expresión genética normalizados en escala $\log_2(x+1)$. Esta transformación es estándar en bioinformatica porque:
- comprime el rango dinámico: permite comparar genes con niveles de expresión masivos (ej. $13-14$ en escala log) frente a genes de baja abundancia (ej.$1-25$).
- normaliza la varianza: mitiga el sesgo de los valores extremos, facilitando la aplicación de modelos lineales y pruebas estadísticas.

A primera vista, de los datos clínicos podemos observar varias cosas:

- La pureza tumoral: Esta podría influir en la obtención de resultados (falsa expresión de genes:

  Un tumor no es una masa aislada de cáncer; es un ecosistema. La pureza disminuye cuando hay una alta presencia de:

  - ***Células del Estroma:*** *Fibroblastos y vasos sanguíneos que el tumor "recluta" para crecer.*
  - ***Células Inmunes:*** *Linfocitos o macrófagos que el cuerpo envía para atacar al tumor o que el tumor manipula para protegerse.*

  La pureza afecta directamente la interpretación de tus resultados:

  - ***Dilución de la señal:*** *Si una muestra tiene solo un 30% de pureza, la señal de los genes que son exclusivos de las células cancerosas será más débil porque está "diluida" por el ARN de las células normales.*
  - ***Falsos positivos/negativos:*** *Algunos genes que marcaste como diferenciales podrían no ser del cáncer en sí, sino de la respuesta del cuerpo (sistema inmune) que aumenta en estadios avanzados.*

En principio observamos que los valores que aparecen como primeros en la tabla son altos (para que os hagais una idea, se considera que el umbral de pureza estandar para valorar la expresión génica de la muestra debe ser superior al 60%). Sin embargo, el mínimo que nos aparecen es 28%, por lo que hay muestras que no cumplen con el umbral. Aquellas que tengan valores nulos también serán eliminadas.

- Sabemos que tanto pathologic_stage/pathology_T_stage como overall_survival/

  overallsurvival son variables que nos dan la misma información, por lo que eliminamos una de ambas

- Si el tipo histologico afecta a lo maligno que sea el tumor (estadío/progresión)

- Pathology_N_stage y pathology_M_stage: ambas describen el proceso de **diseminación**

  (propagación) de las células cancerosas fuera del tumor original. en otras palabras nos indican la progresión del cáncer: si se ha expandido a los ganglios linfáticos o si presenta metástasis

respectivamente, por lo que deben de estar relacionadas (conforme el cáncer progresa, primero se esparce por los ganglios y luego procede a la metástasis)

- Nos surge la pregunta de si podemos observar también correlación entre el estadío del cáncer con estas dos variables que hemos mencionado, o con este y el número de nodos linfáticos afectados (n1)
- Consistencia: ¿Los pacientes clasificados como N1 (afectación ganglionar) muestran la misma firma genética que los pacientes en pathologic_stage III/IV?
- Carga tumoral: ¿Existe una correlación lineal entre el número exacto de ganglios afectados (number_of_lymph_nodes) y el nivel de expresión de tus genes top (COL11A1)?
- También nos parece interesante explorar el overall_survival: Esos números representan el **tiempo** (normalmente en días) desde el diagnóstico hasta el último contacto.
    - Para los que tienen **status 0**, ese número es simplemente cuánto tiempo llevaban vivos la última vez que se tomaron muestras de ellos para el estudio.
    - Para los que tengan **status 1**, ese número indica exactamente cuántos días sobrevivieron
      Nos interesaría quizás hacer una separación para observar los valores de expresión de genes para aquellos pacientes que ya hayan fallecido?
- Hay alguna dependencia entre la progresión de la enfermedad con las características demográficas? (si afecta más a mujeres/hombres, la edad influye en la progresión, etc)
- Existe relación entre si el paciente se encuentra recibiendo radioterapia y el tumor residual?- se supone que esto tengo que verlo en aquellos pacientes a los que no se les haya extendido el tumor???

aumenta el mal pronóstico (o está relacionado algún tipo concreto) dependiendo de la edad?

Hacemos una limpieza de muestras de pacientes, aquellas que no aparecen en ambos dataframe han sido eliminadas, y también hemos filtrado los genes para limpiar el ruido que pueden producir genes que no presentan alta expresión en las muestras (media por debajo de 0.5) , pues no nos interesan, y aquellos que no presenten gran variabilidad de expresión (estamos buscando genes que tengan diferencias observables).- Nos hemos quedado con 14058. También hemos eliminado los valores de pureza tumoral teniendo en cuenta lo que he explicado antes

ANALISIS UNIVARIANTE
Hemos calculado la cardinalidad antes del análisis univariante para ver cómo podemos atacar cada variable, y podemos observar:

attrib_name	Card	%_Card	Clasificada_como
years_to_birth	73	14.570858	Numerica Discreta
Tumor_purity	370	73.852295	Numerica Continua
pathologic_stage	4	0.798403	Categorica
pathology_T_stage	4	0.798403	Categorica
pathology_N_stage	2	0.399202	Binaria
pathology_M_stage	2	0.399202	Binaria
histological_type	4	0.798403	Categorica
number_of_lymph_nodes	29	5.788423	Numerica Discreta
gender	2	0.399202	Binaria
radiation_therapy	2	0.399202	Binaria
residual_tumor	3	0.598802	Categorica
race	4	0.798403	Categorica
ethnicity	2	0.399202	Binaria
overall_survival	444	88.622754	Numerica Continua
status	2	0.399202	Binaria
overallsurvival	449	89.620758	Numerica Continua

attrib_name	Card	%_Card	Clasificada_como	Importancia
edad	73	15.400844	Numerica Discreta	1
pureza_tumoral	351	74.050633	Numerica Continua	0
estadio_patologico	4	0.843882	Categorica	0
estadio_N	2	0.421941	Binaria	1
estadio_M	2	0.421941	Binaria	1
tipo_histologico	4	0.843882	Categorica	0
num_ganglios_linfaticos	26	5.485232	Numerica Discreta	1
genero	2	0.421941	Binaria	2
radioterapia	2	0.421941	Binaria	1
tumor_residual	3	0.632911	Categorica	1
raza	4	0.843882	Categorica	2
etnia	2	0.421941	Binaria	2
supervivencia_global_dias	421	88.818565	Numerica Continua	0
estado_vital	2	0.421941	Binaria	0

Qué observamos de las distribuciones categóricas:

- Los estadios muestran que hay mayor cantidad datos recogidos de pacientes del estadio I que para el resto. Antes de agruparlos aún mas (ya que normalmente se agrupan estadio I y II por una parte, reconociendose como estadios tempranos, y III y IV por otra como tardíos) nos gustaria ver si hay mayor parecido molecular entre I y II o entre II y III/IV. En el segundo de los casos, nos vendría mejor la agrupación, pues tendríamos mas o menos la misma distribución de valores para las agrupaciones. (esto lo haremos en análisis bivariante)
- Hay buena distribución entre aquellos cuyos tumores se han expandido a los nodos linfáticos y aquellos que tienen el tumor todavía focalizado en la zona en la que se produjo. Lo mismo nos conviene centrarnos mejor en esta agrupación para el estudio de la expresión de los genes en la progresión de la enfermedad
- Pocos datos de pacientes con metástasis, nos interesaría saber si aquellos pocos para los que hay datos (ya que hay gran cantidad de pacientes para los que no se ha recogido esta información) siguen vivos.

- Además podemos decir que nuestros datos presentan consistencia. Por lo que sabemos de la literatura del cancer de tiroides:
    - la mayoría está en **estadio I** y casi nadie en **estadio M1** (metástasis a distancia).Este tumor suele detectarse pronto y rara vez viaja a órganos lejanos como pulmones o huesos en el momento del diagnóstico inicial. Sin embargo, es muy común que salte a los ganglios del cuello (tu gráfica muestra muchos **N1**)
    - El cáncer de tiroides es entre **3 y 4 veces más frecuente en mujeres** que en hombres a nivel mundial. No ha habido sesgos en la recolección de datos, sino que la mayor cantidad de datos para mujeres reflejan la realidad epidemiológica de la enfermedad.
    - **estado_vital** muestra que casi todos los pacientes están en "0" (vivos). El carcinoma papilar de tiroides tiene un pronóstico excelente, con una tasa de supervivencia a 10 años superior al **90-95%**. Nos interesaría, como hemos dicho antes, observar las diferentes variables de aquellos pacientes fallecidos para ver si hay alguna marca importante que implique mayor probabilidad del suceso, además de observar la marca molecular de cada uno de dichos pacientes.
    - El hecho de que el tipo **Clásico** sea el predominante frente al **Tall Cell** es lo esperado. El tipo clásico representa el grueso de los diagnósticos, mientras que las variantes como "Tall Cell" son menos frecuentes pero más agresivas. Nos gustaría observas si realmente podemos observar esto en los datos.
- Hubiese sido interesante estudiar la recurrencia de la enfermedad, pero no han dado datos de esta.

Con respecto a las variables numéricas: BOXPLOT

- Hay una buena representación de todas las edades
- Las muestras presentan, en general una pureza tumoral muy alta. Como previamente ya hemos definido el umbral al 60% no nos desharemos de los outliers. Nos gustaría aún así tenerlos en cuenta para ver si tiene algo que ver con la progresión de la enfermedad, ya que cuanto más desarrollado esté el tumor, más presencia de células del estromas y sistema inmune puede presentar.
- Separaremos los outliers de el numero de gánglios linfáticos: **Interés en casos extremos:** nuestro objetivo es encontrar genes que expliquen por qué algunos pacientes tienen metástasis ganglionares masivas, los outliers son los pacientes más interesantes.
- la supervivencia global no nos dice gran cosa, puesto que no sabemos si el paciente ha remitido, además de que los valores pequeños no indican que han sobrevivido más días, sino que se han incorporado más tarde al estudio. podríamos evaluar si aquellos pacientes con valores outliers con mayor supervivencia tienen en su registro un mejor pronóstico.

MEDIDAS DISPERSIÓN.

- Edad mínima 15 años (estudiar esos casos?)
- Como hemos visto el número de gabgios, la mediana es 1, la gran mayoria de los datos o no presentan expansión a los nodos linfáticos o sólo tienen 1 ganglio afectado.

HISTOGRAMA Y FUNCIÓN DENSIDAD.

- **Dinámica de los Ganglios (number_of_lymph_nodes)**

  El histograma tiene un pico masivo en 0 y cae drásticamente. Sin embargo, el boxplot muestra una enorme cantidad de **outliers** que llegan hasta los 40 ganglios. La mayoría de los pacientes no tienen ganglios afectados o tienen muy pocos, pero existe un subgrupo pequeño con una **enfermedad linfática muy agresiva**. Vamos a dividir a los pacientes para estudiarlos en función del numero de ganglios linfáticos en 0 (no presentan), entre 0 y 5 y de 5 a 40. Esto no se corresponde a la separación con los outliers de los boxplot, pero creemos que representará mejor los diferentes perfiles a evaluar.

- **Sesgo en la Supervivencia (overall_survival)**

  Como de todos los datos no sacamos gran información, estudiaremos la supervivencia global para aquellos pacientes que ya están clasificados como fallecidos (ahí, la supervivencia global si que nos dice cuántos dias pasaron desde el diagnóstico hasta la muerte).

Con respecto a la expresión de los genes, vemos Boxplots y distribución, y podemos observar que, tras el tratamiento y la limpieza de datos, no hay ningún tipo de “ruido”. No encontramos valores de outliers, y la curva de distribución tiene una forma de "campana" casi perfecta (distribución normal), lo que indica que el proceso de secuenciación y normalización fue muy consistente entre todos los pacientes. Tras el limpiado no hay muestras "extrañas" que tengan muchísima o poquísima expresión respecto a las demás. Todas las muestras son comparables entre sí.

ANÁLISIS BIVARIANTE.

En general, queremos estudiar esto:

Variable Principal	**Comparada con...**	**¿Por qué? (Objetivo del Análisis)**
estadio_patologico	estadio_N y estadio_M	Validar la consistencia clínica: ¿Se corresponde el avance del estadio con la propagación a ganglios y metástasis?[5].

tipo_histologico	estadio_patologico	Comprobar si variantes como "Tall Cell" presentan mayor agresividad (estadios más avanzados/mayor diseminación) que el tipo "Clásico"[6666].
	estadios n y m	
pureza_tumoral	estadio_patologico	Explorar si los tumores más avanzados presentan menor pureza debido al reclutamiento de células del estroma y sistema inmune[7].
edad	estadio_patologico	Analizar si el diagnóstico en edades avanzadas está relacionado con un aumento del mal pronóstico o estadios más graves[888].
radioterapia	tumor_residual	Estudiar la relación entre el tratamiento recibido y la presencia de restos tumorales tras la intervención[9].

supervivencia_global_dias	estado_vital (Fallecidos)	Evaluar el tiempo real de supervivencia desde el diagnóstico hasta la muerte para caracterizar a los pacientes de mayor riesgo[10].
genero	estadio_patologico	Investigar si la progresión de la enfermedad presenta diferencias significativas entre hombres y mujeres[11].
numero de ganglios	estadio patologico, tipo histológico	
	edad	

Expresión Génica **(esto lo haremos en multivariante)**	estadio_patologico (I/II vs III/IV)	Determinar si existe una firma molecular que diferencie claramente los estadios tempranos de los tardíos.

	perfil_ganglionar (0, 1-5, >5)	Investigar si el nivel de expresión de genes top (como *COL11A1* o *MMP13*) correlaciona con la carga de enfermedad linfática.
	tumor_residual	Evaluar si hay una diferencia en la firma genética entre pacientes que terminaron sin tumor residual tras la cirugía frente a los que sí.
	estado_vital (Fallecidos)	Identificar marcas moleculares específicas en pacientes fallecidos que puedan servir como predictores de mal pronóstico.

Hemos empezado por el análisis donde las dos variables en estudio son categóricas.

1. **Comparación del Estadío con la afectación ganglionar:** Podemos observar como si que hay una relación entre el estadio y la afectación, observando un mayor porcentaje de expansión a los nodos linfáticos en estadíos más avanzados. Esto, en realidad, no había que comprobarlo ya que en realidad los estadíos son categorizaciones que utilizan los médicos para agrupar a los pacientes en grupos en función de como de avanzada esté la enfermedad.

2. **Con la metástasis ocurre lo mismo**, la presentan grupos más avanzados, pero nunca una persona clasificada como en el estadío I. Normalmente, en la mayoría de canceres, la metástasis está directamente relacionada con personas en el estadio IV. En este caso, hay personas en el estadío II (3 en particular) que ya lo presentan. Buscando en la literatura de este cáncer descubrimos que una persona con cáncer de tiroides en estadio 2 puede presentar metástasis a distancia, ya que la Etapa II, se define precisamente por la propagación a órganos distantes como pulmones o huesos, aunque en la mayoría de los casos es infrecuente.

3. **Tipo_histologico vs estadio:**

   - El gráfico de barras confirma visualmente lo que sospechábamos: Mientras que el tipo **Clásico** tiene una gran base en el **Estadio I** (barra azul dominante), el tipo **Tall Cell** muestra un comportamiento radicalmente distinto: su barra más alta corresponde al **Estadio III** (barra verde). Esto demuestra que la variante *Tall Cell* tiende a diagnosticarse en estadios mucho más avanzados y es intrínsecamente más agresiva que la variante clásica

   -Variante Folicular: Muestra un perfil similar al clásico pero con una presencia algo mayor de Estadio II. Podría ser interesante ver si su firma genética es un "punto medio" entre el clásico y los más agresivos.

El p-valor obtenido es extremadamente bajo. Es decir, la relación que ves entre el tipo de tumor y el estadio no es casualidad. Hay una **dependencia biológica real** entre la histología del cáncer de tiroides y su capacidad de propagación. Sin embargo, hay que tener en cuenta que **Categoría "Otros":** Solo tienes datos para el Estadio I y IV. Al ser una muestra tan pequeña (tabla de frecuencias esperadas con valores menores a 5), los resultados para este grupo específico deben tomarse con cautela El enorme volumen de datos del Estadio I puede enmascarar lo que ocurre en los estadios tardíos.

Los resultados "encajan" con la realidad epidemiológica: el cáncer de tiroides suele detectarse en etapas tempranas (Estadio I) y el tipo Papilar Clásico es el más frecuente y de mejor pronóstico. El dataset es una representación fiel de la enfermedad, lo que da mucha validez a nuestros futuros hallazgos genómicos.

4. **Tipo histologico vs estadio_N:** En el tipo **Clásico**, hay más pacientes con afectación ganglionar (**n1**: 168) que sin ella (**n0**: 141).Esto refuerza que es muy que este cáncer salte a los ganglios del cuello incluso en sus variantes menos agresivas. El tipo **Tall Cell** mantiene esta tendencia, teniendo casi el doble de casos con ganglios afectados (**n1**: 19) que sin ellos (**n0**: 12).

5. **Tipo histologico vs Estadio M:** La metástasis distante ($M1$) es extremadamente inusual en toda la cohorte, validando que la progresión es principalmente regional. Que el gráfico muestre 0 casos M1 para tall cell responde principalmente a una limitación de la muestra: el tamaño del subgrupo Tall Cell es muy pequeño comparado con el Clásico, y la metástasis a distancia es un evento extremadamente raro en tiroides que suele detectarse de forma tardía. Además, ya se advertía sobre la falta de datos registrados para la variable metástasis en muchos pacientes, lo que sugiere que la agresividad del Tall Cell en el dataset queda mejor reflejada en su clara tendencia al **Estadio III** y a la afectación ganglionar (**N1**).

6. **Estadio_N vs estadio_M:** En principio esperaríamos que todos los valores de m1 se presentaran en n1, ya que se intuye que la progresión del cáncer comienza siempre por la diseminación linfática y luego aparece la metástasis. Sin embargo, aunque la mayoría de los carcinomas de tiroides prefieren la vía linfática (hacia los ganglios) , el cáncer también puede viajar a través de la **sangre** (vía hematógena).

   Si las células cancerosas entran directamente en un vaso sanguíneo, pueden colonizar el pulmón o el hueso (**m1**) saltándose por completo la estación de los ganglios del cuello (**n0**).

   El caso es que en este caso, es casi igual de frecuente que la metástasis se presente con o sin diseminación linfática. Esto puede deberse a dos casos:

   - Podría ocurrir que el paciente tenga micrometástasis en ganglios que no fueron extirpados en la cirugía, por lo que se registra oficialmente como **n0** (no se encontraron ganglios positivos en la muestra analizada) a pesar de tener ya enfermedad a distancia (**m1**).
   - O al error debido a la baja representación de pacientes con metástasis. No se puede afirmar nada con una muestra tan pequeña.

7. Relación entre **estadio_patologico** **vs** **genero** aporta un dato clínico muy relevante:

- **Prevalencia Femenina**: Se confirma visualmente que el cáncer de tiroides es significativamente más frecuente en mujeres en todos los estadios.
- **Diferencia en Progresión**: El p-valor de **0.0349** indica una relación significativa. Aunque hay más mujeres en total, la proporción de hombres parece aumentar ligeramente en el **Estadio IV** en comparación con el Estadio I, lo que podría sugerir que, aunque los hombres se diagnostican menos, podrían presentar estadios algo más avanzados.

8. Sin embargo, no hay relación con el tipo histologico.

9. Relación entre radioterapia y tumor residual:A simple vista, parece que la radioterapia "empeora" el resultado porque hay más casos de tumor residual (r1). Sin embargo, en oncología clínica, esto suele interpretarse al revés:

   Interpretación probable: Los médicos suelen recetar radioterapia precisamente a los pacientes que tienen **tumores más agresivos, más grandes o más difíciles de extirpar**. Por lo tanto, es normal que en el grupo de "Sí radioterapia" haya más casos de tumor residual, no porque la terapia falle, sino porque esos pacientes ya tenían un pronóstico más complejo desde el inicio.

10. Relación entre estadio patológico y perfil ganglionar:La gráfica confirma que el **Estadio IV** está fuertemente ligado a una **alta carga ganglionar (>5 ganglios)**, lo cual es un indicador de que el cáncer se ha extendido significativamente por el sistema linfático. Por el contrario, en el **Estadio I**, es mucho más probable encontrar pacientes cuyos ganglios están limpios. **Dato curioso:** En el Estadio II hay un gran porcentaje de datos "Desconocidos" (rojo). Esto podría indicar una falta de registros o que en ese estadio específico no siempre se realiza el conteo ganglionar completo.

11. Relación entre tipo histologico y perfil ganglionar: **Propagación Diferenciada:** El tipo **Folicular** es notablemente más localizado (menos propenso a afectar ganglios), mientras que la categoría **Otros** muestra una clara tendencia a la diseminación linfática. **Importancia del Diagnóstico:** Esta gráfica sugiere que conocer el tipo histológico es fundamental para predecir si el paciente necesitará un tratamiento más agresivo en los ganglios.

VARIABLE CATEGÓRICA Y NUMÉRICA.

1. Boxplots of edad for estadio patologico (Group 1): muestra la distribución de la **edad** de los pacientes según su **estadio patológico** (del I al IV). Es una herramienta excelente para ver promedios y variabilidad. La gráfica sugiere una tendencia clara: el diagnóstico en **estadios tempranos (I)** ocurre con mayor frecuencia en personas **jóvenes**, mientras que los **estadios más avanzados (II, III y IV)** se presentan mayoritariamente en adultos de **mediana y tercera edad**.

Conclusión de tus tres gráficas: Si unimos esta información con las anteriores, vemos un perfil de riesgo: los pacientes de mayor edad suelen presentar estadios más avanzados (II-IV), los cuales a su vez tienen una mayor probabilidad de afectación ganglionar agresiva.

2. Distribucion de edad por estadio patologico:**Desplazamiento a la derecha:** Se observa claramente que a medida que el estadio avanza (del I al IV), la distribución de la edad se desplaza hacia la derecha. Esto confirma que los **estadios más graves tienden a diagnosticarse en pacientes de mayor edad.**

   -Solapamiento: Existe un área común entre los 45 y 65 años donde coinciden todos los estadios, lo que significa que en esa franja de edad es posible encontrar cualquier nivel de gravedad de la enfermedad. Rigor Estadístico: En la parte inferior se menciona el **Test de -Kruskal-Wallis**. Este es un test estadístico utilizado para determinar si existen diferencias significativas entre las medianas de estos grupos. Dado que las curvas están tan separadas (especialmente la azul de las demás), este test confirmaría matemáticamente que **la edad sí varía de forma significativa según el estadio patológico**.

3. Boxplots of edad for de progresión binario (Group 1):Al observar todos tus gráficos en conjunto, se consolida un patrón clínico claro:

   - Juventud y Estadios Iniciales: Los pacientes jóvenes (aprox. 37-40 años) suelen estar vinculados al **Estadio I**, a una progresión **Temprana** y a tipos histológicos como el - - -**Folicular**, que tienen menor afectación de ganglios.

   -Edad y Gravedad: Los pacientes de mayor edad (55+ años) tienen una mayor frecuencia de **Estadios II, III y IV**, una progresión **Tardía** y perfiles ganglionares más agresivos (especialmente en Estadio IV).

4. Distribución de edad por progresión binaria (Group 1):Existe un **solapamiento importante** entre los 45 y los 65 años. En este rango de edad, es casi igual de probable pertenecer a cualquiera de los dos grupos, aunque la tendencia de la curva naranja (Tardía) se mantiene más estable en edades avanzadas que la azul.

   En la parte inferior se menciona nuevamente el **Test de Kruskal-Wallis**. Al igual que en las gráficas anteriores, este test se utiliza para confirmar que la diferencia de edad entre el grupo "Temprano" y "Tardío" no es producto del azar, sino que es **estadísticamente significativa**.

5. Boxplots os edad for tipo histologico (Group 1):La gráfica sugiere que el tipo histologico **Tall Cell** está más asociado a pacientes de edad madura, mientras que la categoría **Otros** identifica tumores que ocurren con mayor frecuencia en etapas más tempranas de la vida. Los tipos **Clásico** y **Folicular** se mantienen en un punto intermedio, afectando a un rango de edad muy diverso.

6. Distribución de edad por tipo histologico (Group 1):A través de todas las gráficas analizadas, los datos sugieren que:

   La edad es un factor determinante: A mayor edad, aumenta la probabilidad de diagnósticos de tipo *Tall Cell*, estadios avanzados y progresión tardía.

   Perfiles de Diagnóstico: Los pacientes jóvenes tienden a presentar el estadio I, progresión temprana y tumores con menor propagación ganglionar.

   Significancia Estadística: Los tests de Kruskal-Wallis incluidos en tus imágenes confirman que estas diferencias de edad entre grupos no son casualidad, sino patrones reales en tu base de datos.

7. Boxplots of edad for perfil ganglionar (Group 1): Después de analizar todas las gráficas que has compartido, se pueden extraer estas conclusiones definitivas sobre la base de datos: **Relación Edad-Gravedad:** Existe una correlación clara donde el **Estadio I** y la **progresión temprana** se concentran en pacientes jóvenes (medianas de 37-39 años). En contraste, los **estadios avanzados (II, III, IV)** y la **progresión tardía** afectan principalmente a adultos de más de 55 años.

   Perfil Histológico: El tipo **Tall Cell** es el más asociado a edades avanzadas (mediana >50 años), mientras que el tipo **Folicular** muestra una mayor tasa de éxito sin afectación ganglionar.

   Hallazgo Ganglionar: El **Estadio IV** es el que presenta la mayor frecuencia de enfermedad agresiva (>5 ganglios afectados). Sin embargo, la edad por sí sola no parece ser un predictor único de la carga ganglionar, ya que las medianas de edad entre los grupos de afectación (0, 1-5 y >5) están relativamente cerca entre sí (entre los 43 y 49 años).

8. Boxplots of pureza timoral for estadio patologico (Group 1): La pureza tumoral es un factor crítico en estudios genómicos. Una pureza menor (como se ve en algunos casos del Estadio I o IV) podría dificultar la detección de ciertas mutaciones debido a la "contaminación" con células normales.

9. Boxplots of supervivencia global dias for estadio patologico (Group 1): Existe una clara división por edad. Los **jóvenes** suelen presentar estadios iniciales (I) con tumores tipo Folicular y baja afectación de ganglios. Los **mayores** (55+) tienden a estadios avanzados (III-IV) y tipos como *Tall Cell*. La radioterapia se asocia con grupos que terminan teniendo más tumor residual (probablemente casos más complejos de base). El Estadio II ofrece la mayor pureza tumoral para análisis genéticos. Los estadios avanzados muestran una incertidumbre mucho mayor en los días de supervivencia global comparado con la estabilidad del Estadio I.

ANÁLISIS MULTIVARIANTE:

1. Primero, para comprobar la relación entre la edad con la categorización del estadio en función de la diseminación hemos realizado un bubble plot, donde el tamaño de los puntos representa la cantidad de gánglios afectados. Podemos observar com a edades tempranas, pacientes con gran cantidad de ganglios afectados siguen siendo categorizados como de estadío I. Solo hay un caso para el que esto no se cumple. Sin embargo, a edades más avanzadas, se observa como la clasificación sigue la estructura natural de todos los cánceres (los que presentan menor diseminación en los estadíos primarios y conforme progresa la enfermedad en tardíos).

2. Luego hemos realizado un Volcano plot comparando la expresión diferencial de genes en estadios tardíos vs tempranos. El gráfico tiene líneas punteadas que definen qué genes son "importantes":La línea horizontal: Es el umbral significativo (p-valor de 0.001). Todo lo que esté por encima es estadísticamente significativo. Las líneas verticales: Marcan el umbral de magnitud del cambio.

   Mientras los puntos Grises son genes que no cambiaron lo suficiente su perfil de expresión o cuyos cambios no son estadísticamente confiables, los Puntos Rojos son los genes de interés. Han pasado ambos filtros: cambian mucho y son significativos. Lado derecho (ej. COL11A1, MMP13, CCL17): Estos genes están sobreexpresión en los estadios avanzados. Podrían ser biomarcadores de progresión o responsables de que la enfermedad empeore. Estos genes se pierden o se "apagan" a medida que la enfermedad avanza- Hemos obtenido 39 genes.

3. Ahora, vamos a intentar ver si podemos obtener un claro mapa diferenciativo donde observemos claramente la diferencia de expresión comparando perfiles.

    1. Primero lo comparamos en función del estadío (tanto agrupado, en la primera diapositiva, como sin agrupar). Podemos observar que agrupados encontramos una clara diferencia de patrones. De los 39 genes, algunos en mayor medida que otros, están: 37 sobreexpresados en estadíos tardíos y 2 subexpresados (los dos últimos, se puede notar el cambio de color). Aún así, podemos ver como ciertas muestras no presentan del todo el perfil apagado esperado en los estadíos tempranos, lo que puede deberse justamente a la clasificación en estadíos dependiente de la edad. De hecho, si observamos la firma genética sin agrupar, nos damos cuenta de que el estadío II tiene claramente un perfil más apagado, y esto ya lo podíamos intuir antes con el bubble plot, ya que en este estadío solo encontrábamos pacientes con menor número de ganglios (menor progresión).

    2. Por otro lado, si estudiamos la expresión génica en función de el tipo histológico, también encontramos una clara diferencia de expresión en el tipo folicular, lo que tampoco nos sorprende, porque recapitulando, hemos visto que esta variante presentaba muchísima menor afectación ganglionar.

    3. Todo esto nos lleva a la misma consideración: El análisis de expresión génica más diferenciativo y donde se observará una mejor clasificación será aquel que se estudie en función de la diseminación ganglionar. Se observa en el heatmap una clara diferenciación. Esto pone más en duda esa dependencia de la clasificación por rango de edad, pues hemos visto claramente que el perfil genético asocia sus cambios de expresión al perfil ganglionar.

    4. Vamos, por último, a ver si claramente podemos observar una diferencia de expresión en función del perfil ganglionar, pero esta vez separando entre pacientes mayores y menores de 55 años (el corte de edad al que nos referíamos cuando explicamos todo aquello de la diferente clasificación en estadíos). Se puede observar:

        En el grupo de menores de 55 años:

        - **Señal "Diluida"**: Aunque existe un bloque de expresión roja (alta agresividad) en los pacientes con metástasis ganglionar (**n1**, barra azul), este patrón es más **heterogéneo**. Hay muchos pacientes n1 que todavía muestran perfiles "fríos" (azules).
        - **Infiltrados moleculares**: Se aprecian varios pacientes con estadio n0 (barra roja) que tienen una firma genética casi idéntica a los n1. Esto sugiere que en jóvenes, la maquinaria genética de invasión puede encenderse mucho antes de que se detecte el ganglio físicamente.

        En el grupo de mayores de 55 años:

        - **Bloques más compactos**: El bloque de pacientes n1 presenta una señal roja mucho más **densa y uniforme**.
        - **Separación más clara**: La división entre el perfil "tranquilo" (n0) y el "invasivo" (n1) parece estar más definida visualmente, lo que correlaciona con el hecho de que en mayores la clínica es un predictor más fiable del riesgo que en jóvenes

Vamos a realizar de nuevo el volcano plot, esta vez en función de la presencia o no de ganglios, a ver si esta vez cambia el asunto (ya que los genes que estamos estudiando los hemos sacado comparando el estadiaje e incluso así hemos podido observar en el heatmap que es mejor la separación en función del perfil ganglionar).

Hemos podido observar una mayor cantidad de genes diferencialmente expresados, por lo que hemos aumentado el umbral de magnitud de cambio a 2. De esta forma, el heatmap que obtendremos será incluso más robusto. En este caso obtenemos 27 genes.

Esta vez solo vamos a plantear el heatmap en función del perfil (separando en edades de nuevo. Observamos que, a pesar de que los cambios sean más observables en personas mayores de 55, encontramos que para esta clasificación si hay una diferencia de expresión con más lógica en jóvenes. Por ello, concluimos que estos genes tienen mayor confianza. Sigue habiendo sobre todo ciertos perfiles de jóvenes que tienen un mayor parecido al cáncer diseminado, sería interesante estudiarlos por separado. Además, a pesar de que en la distribución de edad con respecto al perfil ganglionar hayamos visto que hay un mayor porcentaje de jóvenes que presenten un mayor número de ganglios afectados, parece verse que no hay muestras que tengan una diferencia de expresión más alta que otras. El simple hecho de que el cáncer se disemine ya activa/desactiva dichos genes.