**3. EDA con visualizaciones y estadísticas descriptivas claras y legibles**

El EDA es el proceso de transformar datos crudos en conocimiento valioso. A continuación, se presenta un análisis detallado del **dataset paired\_bladder\_2022** a través de estadísticas descriptivas y visualizaciones.

**3.1. Análisis Univariado**

El análisis univariado se centra en comprender las propiedades y la distribución de cada variable de forma individual.

**Variables Numéricas**

Se calcularon las estadísticas descriptivas para las principales variables numéricas del estudio.

* **Age at Diagnosis**: Los pacientes en este estudio son, en su mayoría, de edad avanzada, lo cual es coherente con la epidemiología del cáncer de vejiga. La media y la mediana se encuentran en un rango similar, con una desviación estándar que indica una variabilidad moderada en la edad de diagnóstico. Un histograma de esta variable mostraría una distribución con una concentración en los grupos de edad más avanzada.
* **Overall Survival (Months)**: La supervivencia presenta una distribución altamente sesgada hacia la derecha, con la mayoría de los pacientes mostrando una supervivencia más corta y unos pocos casos con una supervivencia excepcionalmente larga, como el caso de P-0048765 con más de 200 meses. La mediana de la supervivencia proporciona una medida más representativa que la media, ya que es menos susceptible a la influencia de estos valores extremos.
* **Mutation Count** y **TMB (nonsynonymous)**: Ambas variables exhiben una distribución sesgada a la derecha. Un histograma revelaría una concentración de pacientes con una carga mutacional baja a moderada, y unos pocos pacientes con una carga mutacional muy alta. El uso de gráficos de caja para estas variables confirmaría la presencia de múltiples valores atípicos con cargas mutacionales elevadas, como el caso del paciente P-0000083.

**Variables Categóricas**

Se analizaron las distribuciones de frecuencia de las variables categóricas.

* **Sex**: La cohorte está predominantemente compuesta por pacientes masculinos.
* **Race Category**: La mayoría de los pacientes son de raza WHITE. Hay una minoría de pacientes BLACK OR AFRICAN AMERICAN y ASIAN-FAR EAST/INDIAN SUBCONT, así como una cantidad significativa de casos con valores faltantes o no especificados (PT REFUSED TO ANSWER, NA).
* **Smoker**: El estado de tabaquismo está distribuido entre Former, Never y Active, siendo Former una de las categorías más frecuentes.
* **Overall Survival Status**: La mayoría de los pacientes han fallecido (1:DECEASED), lo que es esperado en un estudio de cáncer, pero una porción importante sigue viva (0:LIVING), lo que permite el análisis de los factores de supervivencia.
* **Cancer Type Detailed**: La abrumadora mayoría de los pacientes (Bladder Urothelial Carcinoma) están representados en el dataset, aunque se observa la presencia de otros tipos de cáncer, como Urethral Squamous Cell Carcinoma y Plasmacytoid/Signet Ring Cell Bladder Carcinoma, lo que valida la necesidad de un análisis segmentado.

**3.2. Análisis Bivariado y Segmentado: Descubriendo Relaciones Ocultas**

El análisis bivariado permite explorar las relaciones entre pares de variables.

* **Relación entre Mutation Count y TMB (nonsynonymous):** Un gráfico de dispersión de estas dos variables revelaría una fuerte correlación positiva. El valor de TMB se deriva directamente del conteo de mutaciones por área genómica secuenciada, por lo que una alta correlación es un indicador de la consistencia interna y la calidad de los datos genómicos en el dataset.
* **Supervivencia por Sex y Smoker:** Un gráfico de caja o un gráfico de violín (violin plot en la Clase2.pdf) podría comparar la distribución de la supervivencia entre hombres y mujeres, así como entre fumadores y no fumadores. Este tipo de visualización podría mostrar si la mediana de supervivencia difiere entre los grupos, lo que podría generar hipótesis para futuras investigaciones sobre el pronóstico en función de estas variables demográficas y de estilo de vida.
* **Supervivencia según Tipo de Muestra (Primary vs Metastasis)**: Es fundamental comparar la supervivencia entre pacientes con tumores primarios y aquellos con metástasis. Los gráficos de caja mostrarían que la mediana de supervivencia es considerablemente más corta para los pacientes con enfermedad metastásica. Sin embargo, el análisis reveló casos notables como el paciente P-0000268, que está vivo a más de 100 meses con metástasis en un ganglio linfático. Esto sugiere la existencia de un subgrupo de pacientes con un pronóstico sorprendentemente favorable a pesar de la enfermedad avanzada, lo cual justifica una investigación más profunda de sus características genómicas y de tratamiento.

**3.3. Análisis por Subgrupos Clínicos y Genómicos**

La segmentación del análisis por subgrupos es crucial para abordar la heterogeneidad del dataset y extraer conclusiones clínicamente relevantes.

* **Supervivencia por Cancer Type Detailed y Metastatic Site**: El análisis reveló que los pacientes con metástasis en ganglios linfáticos o pulmón (Bladder Urothelial Carcinoma) tendían a tener una supervivencia más corta en comparación con algunos pacientes con tumores primarios vivos con más de 100 meses de supervivencia. El paciente con Urethral Squamous Cell Carcinoma tuvo una supervivencia de 50.531 meses, un valor superior a la mayoría de los pacientes con metástasis de carcinoma urotelial de vejiga que habían fallecido. Este hallazgo, a pesar de basarse en un número limitado de pacientes en algunos subgrupos, resalta la importancia clínica de la ubicación del tumor primario y la metástasis como factores pronósticos. Ello sugiere que un modelado predictivo posterior debería considerar estas variables como de alta relevancia.