MÁSTER EN DATA SCIENCE & IA

BIG DATA Y SALUD: DATOS MASIVOS AL SERVICIO DE LA PREVENCIÓN DEL CÁNCER

TYSSA MARTÍN FERNÁNDEZ

EVOLVE ACADEMY

ABRIL 2025

ÍNDICE

1. Introducción

El cáncer es una de las principales causas de morbilidad y mortalidad en todo el mundo. Comprender su distribución por tipo, región, sexo y etnia resulta fundamental para diseñar políticas públicas, asignar recursos sanitarios y avanzar en la prevención y tratamiento de la enfermedad. En este proyecto realiza un análisis exploratorio y descriptivo de la base de datos *"Cancer Incidence in Five Continents",* una fuente de datos epidemiológicos recopilada por la Agencia Internacional para la Investigación del Cáncer (IARC).

El objetivo principal es explorar patrones relevantes en la incidencia del cáncer mediante técnicas de ciencia de datos, con especial énfasis en la limpieza, visualización y análisis estadístico de los datos. También se contempla el cálculo de indicadores clave, como la tasa de incidencia por cada 100.000 personas-año.

2. Metodología

El análisis se ha desarrollado en el entorno de programación Python, utilizando el editor Visual Studio Code y un entorno virtual para garantizar la reproducibilidad y el aislamiento de dependencias.

El dataset contiene las siguientes variables:

- id_code: código de la región o país
- sex: sexo de la población (1 = male; 2 = female)
- cancer_code: código del tipo de cáncer
- age: grupo de edades de los 0 a más de 85 años agrupados en 18 grupos
- cases: número de casos registrados, cuando es cero es debido a que el caso ha sido registrado, pero no confirmado científicamente.
- py: *person-years*, es decir, el tiempo acumulado durante el cual la población fue observada
- year: año de observación

Uno de los principales indicadores epidemiológicos calculados es la tasa de incidencia, que permite comparar regiones o poblaciones con diferentes tamaños, calculada dividiendo los casos entre la variable py (person-year) y multiplicando el resultado por 100.000.

3. Integración de datos complementarios

Para enriquecer el análisis y darle mayor profundidad, se han incorporado dos bases de datos adicionales a la base. Estas nuevas fuentes permiten añadir información relevante tanto sobre los tipos de cáncer como sobre las regiones geográficas representadas.

3.1. Diccionario de tipos de cáncer (cancer_dict.csv)

Este archivo contiene información adicional sobre los distintos códigos de cáncer recogidos en la base principal. Cada tipo de cáncer está asociado a un código (cancer_code) que puede referirse a una clasificación general o a un subtipo histológico

más específico. La columna histo_label, en particular, contiene descripciones de los subtipos tumorales. Sin embargo, durante la exploración inicial se detectó que esta columna contiene una proporción significativa de valores nulos. De un total de 170 registros, 60 valores son nulos y 110 contienen descripciones específicas, lo que sugiere que muchos cánceres no tienen un subtipo histológico definido en la base.

3.2. Diccionario de regiones (id_dict.csv)

Este segundo archivo incluye información asociada al identificador id_code, como el continente, la región, la raza principal de la población y otros elementos contextuales relevantes para la interpretación epidemiológica.

3.3. Fusión de las bases de datos

Con el fin de consolidar toda la información en una única estructura de datos, se han realizado dos operaciones de fusión (merge) empleando las claves comunes entre los datasets. El resultado es un único dataframe enriquecido llamado df_final, que contiene tanto la información original de casos, como las descripciones de cáncer y las características regionales.

4. Focalización del análisis: Cánceres del cerebro y sistema nervioso central

Debido a la amplitud del conjunto de datos y a la diversidad de tipos de cáncer representados, opté por centrar el análisis en un subconjunto de especial interés: los tumores del cerebro y del sistema nervioso central. Esta categoría incluye múltiples subtipos clínicamente relevantes, como:

- Tumores astrocíticos
- Tumores oligodendrogliales y gliomas mixtos
- Tumores ependimarios
- Gliomas (otros)
- Meduloblastomas
- Otros tumores embrionarios
- Otros tumores neuroepiteliales
- Morfología especificada
- Morfología no especificada

El resultado fue un subconjunto con 1.366.860 registros relacionados con este grupo de patologías. Sin embargo, una parte significativa de estos registros tenía el valor 'cases' = 0. Esto indica que hay casos registrados, pero no verificados, únicamente sobre la base de un certificado de defunción. Se contabilizaron 1.045.134 registros con cero casos, lo cual representa aproximadamente un 76% del total. Para garantizar que el análisis posterior se basara únicamente en observaciones con presencia real de casos, se eliminaron estas filas.

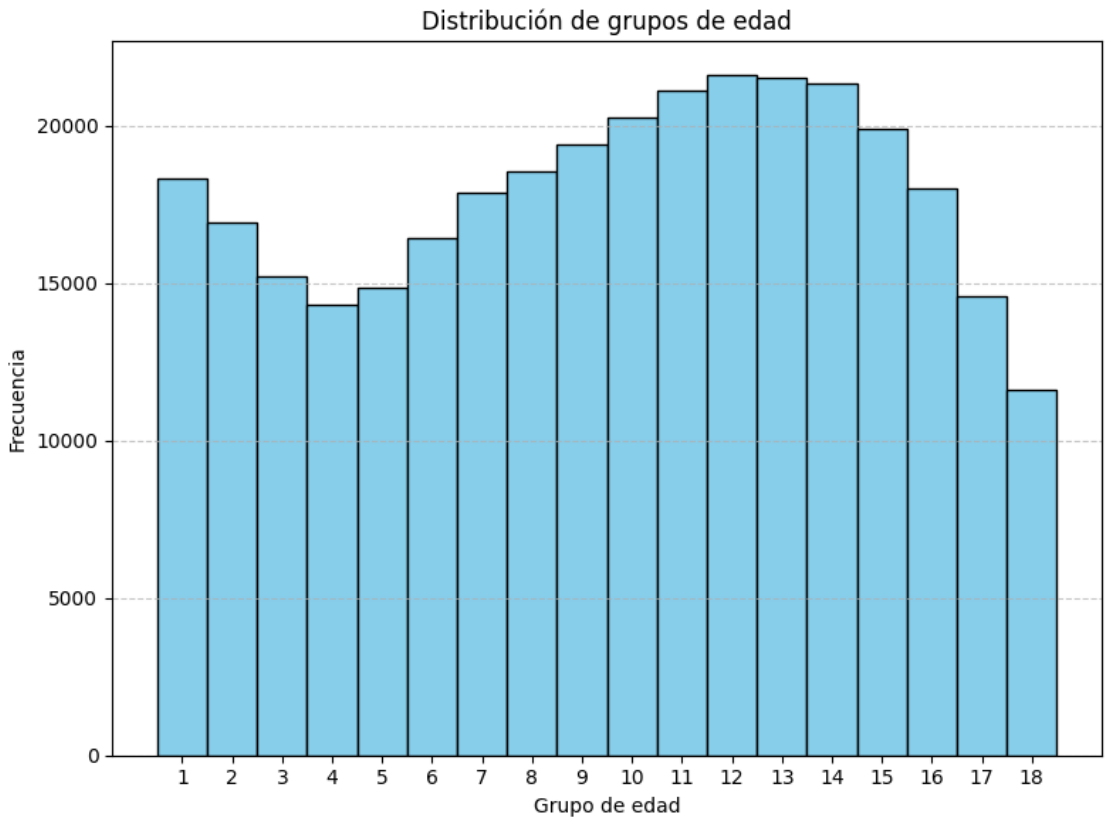
Este filtrado deja una base depurada y más manejable, que refleja solo los registros verificados de cáncer cerebral, y sobre la que se realizarán los análisis estadísticos y visualizaciones en las siguientes secciones.

5. Análisis de distribución por grupos de edad

La variable age no representa edades individuales, sino que agrupa los datos en 18 rangos etarios codificados (como se muestra en la tabla). Para analizar la distribución de los tumores cerebrales según la edad, se ha utilizado esta clasificación y se ha representado gráficamente la frecuencia de casos por grupo.

Age group ID	Age group
1	0–4
2	5–9
3	10–14
4	15–19
5	20–24
6	25–29
7	30–34
8	35–39
9	40–44
10	45–49
11	50–54
12	55–59
13	60–64
14	65–69
15	70–74
16	75–79
17	80–84
18	85+

Tabla 1. Grupos etarios y su correspondiente edad.



Gráfica 1. Histograma por grupo etario en función de la frecuencia de casos de cáncer.

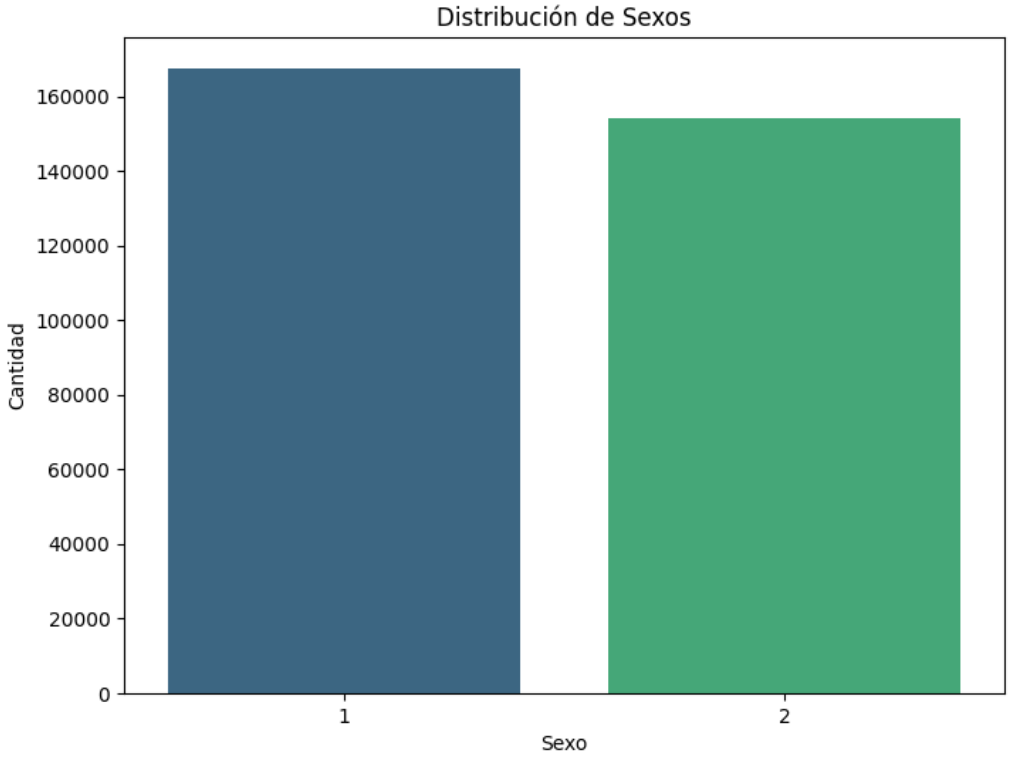
El histograma revela una distribución asimétrica, con un patrón que se puede dividir en tres fases:

- Primera etapa (grupos 1–5): las primeras décadas de vida (infancia y adolescencia temprana, de los 0 a los 24 años), los casos comienzan relativamente altos, pero van disminuyendo.
- Etapa intermedia (grupos 6–13): la adultez media, donde parece concentrarse el mayor número de diagnósticos. Se observa un incremento progresivo en la frecuencia de casos, alcanzando el pico máximo entre los grupos 11 y 13.
- Etapa final (grupos 14–18): tras el pico, hay una disminución sostenida en la frecuencia, lo que indica que la incidencia tiende a reducirse en los grupos de edad más avanzados.

Este comportamiento sugiere que los tumores cerebrales tienden a ser más frecuentes en adultos de mediana edad, con menor prevalencia tanto en edades muy tempranas como en las más avanzadas.

Este tipo de visualización facilita la detección de patrones de incidencia y permite enfocar futuras comparaciones con otras variables, como el sexo, el año o el país.

6. Distribución por sexo



Gráfica 2. Gráfica de barras para determinar la cantidad de casos de cáncer en función del sexo.

La incidencia de tumores cerebrales es ligeramente superior en hombres (1) que en mujeres (2). Se observa una diferencia clara, aunque no extrema, en el número de registros: alrededor de 167.000 casos en el caso de los hombres frente a los 154.000 casos en mujeres.

Este resultado es consistente con diversas investigaciones epidemiológicas que sugieren que algunos tipos de cáncer cerebral, especialmente los gliomas, presentan mayor prevalencia en la población masculina. Esta diferencia puede estar influenciada por factores hormonales, genéticos o ambientales.

7. Cálculo de la tasa de incidencia y evolución temporal

7.1. Tasa de incidencia

Uno de los indicadores epidemiológicos más relevantes para estudios de salud pública es la tasa de incidencia, que permite comparar la frecuencia de aparición de una enfermedad independientemente del tamaño de la población observada. En este caso, la tasa se ha calculado mediante la fórmula estándar:

$$Tasa\ de\ incidencia = (cases/py)\ \times 100.000$$

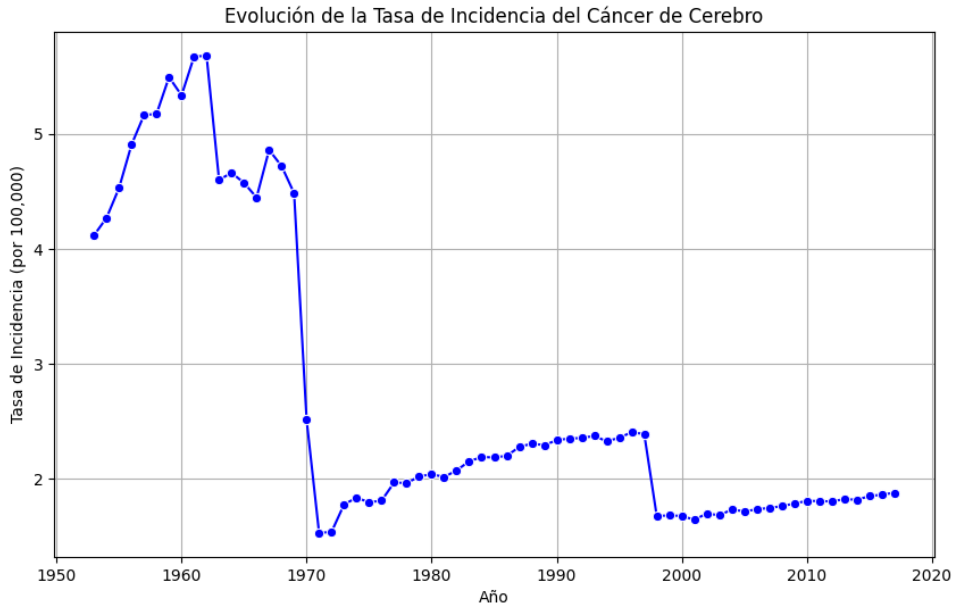
Este cálculo se ha implementado en Python añadiendo una nueva columna incidence_rate al dataset.

7.2. Evolución temporal

Con la tasa de incidencia ya calculada, se ha realizado un análisis temporal para observar cómo ha evolucionado la incidencia de tumores cerebrales a lo largo de los años. Para ello, se ha agrupado el conjunto de datos por año y se ha calculado la media anual de la tasa de incidencia.

Para que fuera más precisa, se eliminaron todas aquellas observaciones en las que la variable py (person-years) era igual a cero o contenía valores nulos, ya que estos registros no permiten calcular una tasa de incidencia válida.

Se recalculó la tasa de incidencia esta vez utilizando el total de casos dividido por el total de *person-years* de cada año, lo cual representa una estimación más sólida a nivel poblacional.



Gráfica 3. Evolución de la tasa media de incidencia desde el año 1953 al 2017.

Fase de aumento progresivo (1953–1965). Durante la primera década, se observa un crecimiento constante de la tasa de incidencia, alcanzando un pico cercano a 5’6 en los años 60. Este ascenso se puede deber a la mejora progresiva de los sistemas de detección y diagnóstico o un aumento en la exposición a factores de riesgo ambientales o laborales no identificados aún.

Descenso brusco (1970–1971). A partir de 1970, se produce un descenso abrupto en la tasa de incidencia, bajando casi dos puntos completos en un año. Este cambio tan marcado puede que no sea una variación real en la enfermedad, sino inconsistencias en el sistema de registro, como por ejemplo la transición de formatos manuales a digitales.

Estabilización y ligera recuperación (1975–1997). En este tramo, la tasa se estabiliza en torno a valores de 1’8 y 2’5 casos. A pesar de ligeras oscilaciones, no se observan picos tan pronunciados como en las décadas anteriores.

Segundo descenso (1998–2000) y estabilización reciente (2000–2017). Entre 1998 y 2000 se identifica otro descenso, tras el cual la tasa se mantiene estable, en torno a los 1'6–1'9 casos hasta el final del periodo analizado.

Este comportamiento sugiere que los valores más recientes son más estables y probablemente más confiables desde el punto de vista estadístico, ya que reflejan una mejora global en los sistemas de vigilancia epidemiológica. Por tanto, las comparaciones por subgrupos (sexo, edad, región) deberían centrarse preferentemente en los datos posteriores a los años 80–90.

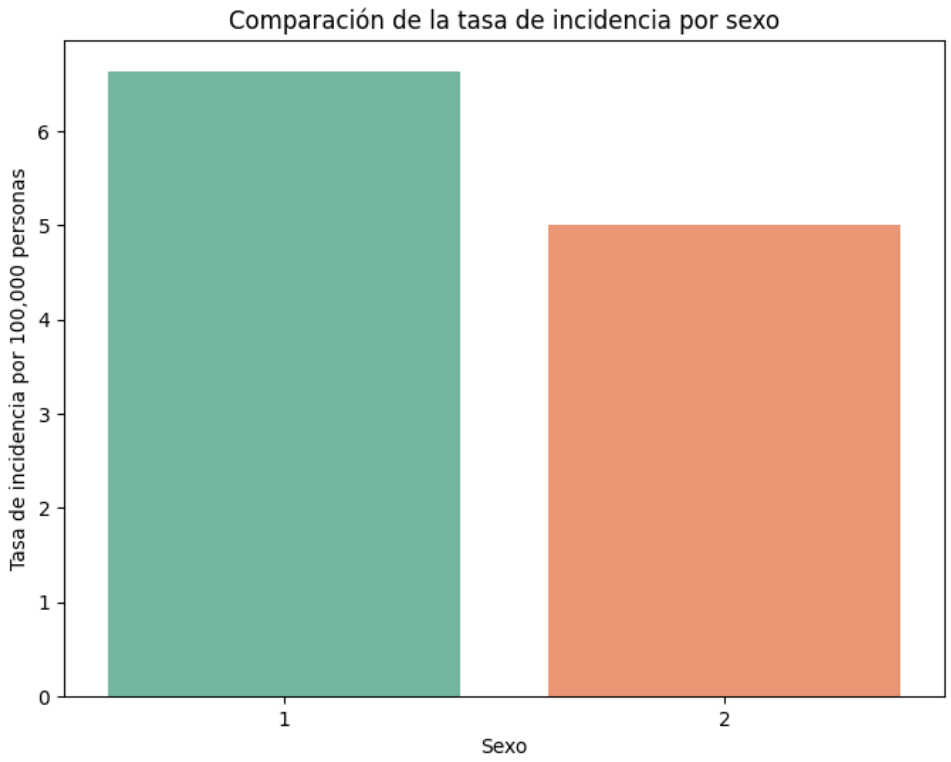
7.3. Tasa de incidencia media por grupo de edad

Gráfica 4. Evolución de la tasa media de incidencia por grupo etario.

El análisis por grupos de edad muestra una tasa de incidencia baja y estable durante la infancia y adolescencia temprana (grupos 1–6), con valores entre 1,5 y 2,1 casos por cada 100.000 personas. A partir de la adultez (grupos 7–13), la incidencia aumenta de forma progresiva, alcanzando cerca de 10 casos, lo que sugiere una mayor vulnerabilidad vinculada a factores biológicos, ambientales y ocupacionales. En la edad avanzada (grupos 14–18), la curva continúa en ascenso hasta su punto máximo en el grupo 17, con más de 15 casos por 100.000 personas, seguido de una ligera disminución en el último grupo, aunque manteniéndose en niveles elevados.

Este patrón de distribución refuerza la idea de que la incidencia del cáncer cerebral está fuertemente asociada a la edad, con una curva ascendente que se acentúa especialmente a partir de los 50 años (grupos 11–12 en adelante). El leve descenso en el grupo 18 podría deberse a una menor detección en personas muy mayores, o menor esperanza de vida que impide el diagnóstico.

7.4. Comparación de la tasa de incidencia por sexo



Gráfica 5. Comparación de la tasa media de incidencia por sexo.

Esta diferencia refleja una mayor incidencia de tumores cerebrales en la población masculina (6’63 por 100.000 personas) frente a la femenina (5 por 100.000 personas).

Aunque la magnitud de la diferencia no es extrema, sí puede tener implicaciones relevantes para la investigación biomédica y el diseño de campañas de prevención o detección precoz con enfoque de género.

8. Estadística descriptiva de la tasa de incidencia

8.1. Medidas de tendencia central

Para comprender la distribución de la tasa de incidencia del cáncer cerebral, se calcularon las principales medidas de tendencia central. La media fue de 5,86 casos por cada 100.000 personas-año, la mediana de 2,46 y la moda de 1,35, esta última con una frecuencia de 19. La diferencia notable entre la media y la mediana sugiere una distribución asimétrica, en la que algunos valores extremos elevan la media respecto al valor central real de los datos.

8.2. Medidas de dispersión

Las siguientes métricas permiten evaluar la variabilidad en las tasas de incidencia: el rango fue de 877,18, la varianza de 86,98 y la desviación estándar de 9,33. Estos resultados confirman una alta dispersión en los valores, lo que refuerza la idea de que existen regiones, edades o situaciones muy específicas con tasas excepcionalmente

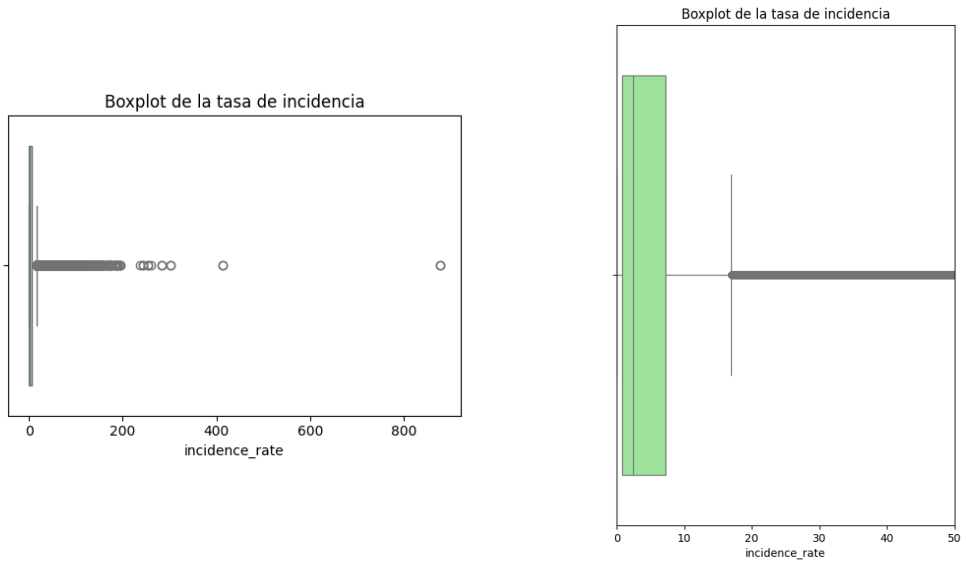
elevadas, aunque la mayoría de los casos se concentran en valores considerablemente más bajos.

8.3. Percentiles

Se calcularon varios percentiles clave para analizar la distribución acumulativa de la tasa de incidencia. El percentil 25 se situó en 0,77 casos por cada 100.000 personas, el percentil 50 (mediana) en 2,46 y el percentil 75 en 7,26. Los valores más elevados muestran que el 90 % de los casos están por debajo de 15,83, el 95 % por debajo de 22,49 y solo un 1 % supera los 39,53. Estos resultados confirman que la mayoría de los valores se concentran en niveles bajos, mientras que una pequeña proporción representa tasas excepcionalmente altas, probablemente asociadas a zonas geográficas concretas o situaciones atípicas.

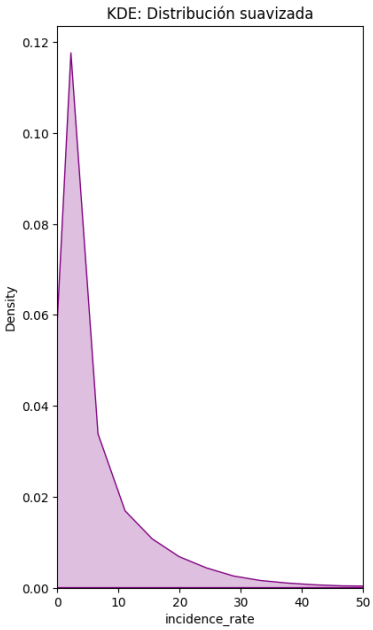
8.4. Boxplot de la tasa de incidencia

La mayoría de las tasas se encuentran concentradas entre 0 y aproximadamente 20 casos por 100.000, como ya anticipaban los percentiles (P75 ≈ 7,26). Sin embargo, hay valores extremos, lo que confirma una distribución altamente asimétrica con sesgo positivo (cola larga a la derecha).Estos outliers podrían corresponder a áreas geográficas pequeñas con tasas inusualmente altas, años específicos con picos, o incluso errores de registro.



Gráfica 6. Boxplot (a la derecha ampliado) para ver posibles outliers en la tasa de incidencia.

8.5. Distribución suavizada (KDE) de la tasa de incidencia



La curva presenta una asimetría marcada hacia la derecha, característica de una distribución sesgada positivamente.

El máximo de densidad se concentra en los 5 casos por cada 100.000 personas, lo que indica que la gran mayoría de los registros tienen tasas bajas.

A partir de ese pico inicial, la curva desciende rápidamente, reflejando que los valores más altos son mucho menos frecuentes, pero todavía presentes.

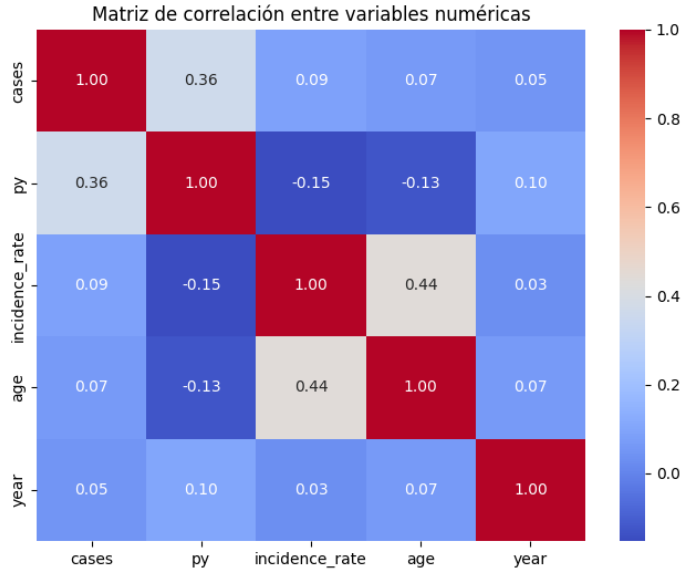
La larga cola hacia la derecha confirma la presencia de valores atípicos o extremos, ya identificados previamente mediante boxplot y percentiles.

Gráfica 7. KDE de la tasa de incidencia.

9. Estadística inferencial

9.1. Análisis de correlación entre variables numéricas

Para explorar posibles relaciones lineales entre las variables numéricas del conjunto de datos, se calculó la matriz de correlación de Pearson.



Gráfica 8. Matriz de correlación entre la tasa de incidencia, la edad, el sexo, el número de caos y los años en los que se registraron.

El análisis de correlaciones revela una relación positiva moderada (0,36) entre las variables *cases* y *py* (person-years), ya que un mayor tiempo de observación suele asociarse a un mayor número de casos registrados. La variable *incidence_rate*, que representa la tasa ajustada por 100.000 personas, no muestra una relación fuerte ni con *cases* ni con *py*, ya que precisamente se construye para normalizar esas diferencias. Se observa, además, una correlación moderada entre *incidence_rate* y *age* (0,44), lo que respalda los análisis previos que indican que la incidencia del cáncer cerebral aumenta con la edad. En cambio, las correlaciones entre *year* y el resto de las variables son muy bajas (todas por debajo de 0,10), lo que sugiere que el año no influye de forma lineal directa. No se detectan correlaciones altas que apunten a multicolinealidad entre potenciales predictores. Por último, aunque la variable *sex* se almacene como numérica, se trata de una variable categórica binaria, por lo que aplicar una correlación de Pearson no refleja una relación lineal real entre el sexo y otras variables cuantitativas.

9.2. Contraste de hipótesis: ¿es diferente la tasa de incidencia entre sexos?

Con el objetivo de comprobar si hay diferencias estadísticamente significativas entre las tasas de incidencia de tumores cerebrales en hombres y mujeres, se aplicó una prueba t de Student para muestras independientes (Welch's t-test, sin asumir varianzas iguales).

Los resultados obtenidos fueron un estadístico t de 49,94 y un p-valor < 0,0001, lo que lleva a rechazar la hipótesis nula de igualdad de medias. Esto significa que la diferencia entre las tasas de incidencia en hombres y mujeres es estadísticamente significativa.

Este resultado confirma lo observado en los análisis descriptivos: los hombres presentan, en promedio, una mayor tasa de incidencia de cáncer cerebral que las mujeres. Dado el elevado valor del estadístico t, se trata de una diferencia marcada y robusta, no atribuible al azar.

9.3. Modelo de regresión lineal: predicción de la tasa de incidencia

Variable	Coeficiente	Error estándar	Valor t	p-valor	Intervalo de confianza (95%)
Constante	-1.1310	2.491	-0.454	0.650	[-6.013 ; 3.752]
Edad	0.8345	0.003	279.268	0.000	[0.829 ; 0.840]
Sexo	-1.8097	0.029	-61.401	0.000	[-1.867 ; -1.752]
Año	0.0008	0.001	0.674	0.500	[-0.002 ; 0.003]

Métrica	Valor
R^2	0.203
R^2 ajustado	0.203
F-statistic	27.140
Probabilidad (F-statistic)	< 0.001
Durbin-Watson	0.960
Jarque-Bera	9.3e+09
Skew	12.416
Kurtosis	838.129
Número de observaciones	320.046

En la variable edad el coeficiente es positivo (0.8345) y altamente significativo ($p < 0.001$), lo que indica que, cada unidad de aumento en el grupo de edad se asocia con un aumento de 0.83 casos por cada 100.000 personas en la tasa de incidencia. Esto es coherente con el análisis descriptivo: la incidencia aumenta con la edad.

En el sexo, el coeficiente es negativo (-1.8097) y muy significativo. Dado que sex está codificado como 1 (hombre) y 2 (mujer), esto indica que, en igualdad de condiciones, las mujeres presentan tasas de incidencia 1.81 puntos más bajas que los hombres, lo que refuerza los hallazgos anteriores sobre diferencias de género.

El coeficiente de year es muy bajo (0.0008) y no es estadísticamente significativo ($p = 0.50$), lo que sugiere que, una vez controladas la edad y el sexo, el año de registro no aporta valor predictivo adicional.

El modelo de regresión lineal resulta globalmente significativo, con un valor de *F-statistic* igual a 27.140 y un valor-p inferior a 0,001. Esto indica que, en conjunto, las variables independientes explican una proporción significativa de la variabilidad observada en la tasa de incidencia. Sin embargo, el estadístico de Durbin-Watson —que evalúa la autocorrelación de los residuos del modelo— tiene un valor de 0,96, lo que sugiere la presencia de cierta autocorrelación positiva. Esto implica que los errores no son completamente independientes entre sí. Por otro lado, la prueba de normalidad de Jarque-Bera, junto con una elevada curtosis (superior a 800), revela que los residuos no siguen una distribución normal. Aunque esto podría afectar la validez de algunas inferencias estadísticas (como los intervalos de confianza o la significación de los coeficientes), es algo esperable en muestras muy grandes y en presencia de variables con distribuciones sesgadas, como se ha observado previamente en el KDE y el boxplot de la tasa de incidencia.

Como conclusión, el modelo de regresión confirma que la edad y el sexo son factores determinantes en la tasa de incidencia del cáncer cerebral, mientras que el año no resulta significativo. Aunque el modelo no explica una proporción muy alta de la variabilidad (R^2 = 0.20), los resultados son coherentes con el análisis descriptivo previo y respaldan la importancia de la edad avanzada y el género masculino como factores de riesgo asociados a una mayor incidencia.

10. Visualización de resultados con Power BI

Para complementar el análisis exploratorio y estadístico realizado en Python, se construyó un dashboard interactivo en Power BI. Se utilizó como fuente de datos el archivo limpio cancer_brain_cleaned.csv, al que se unieron dos tablas adicionales: age_groups (que traduce los códigos de edad en grupos etarios interpretables) y continent (que convierte los códigos CI5_continent en nombres de continentes).

El modelo de datos se organizó mediante relaciones uno a muchos desde las tablas auxiliares hacia la principal:

- La tabla continent se conectó a cancer_brain_cleaned a través de la columna CI5_continent.
- La tabla age_groups se relacionó mediante la columna age.

Estas relaciones permitieron enriquecer las visualizaciones con categorías más comprensibles, como los nombres de continentes o los rangos de edad.

Hubo un problema en los valores de la columna incidence_rate, que aparecían con cifras desproporcionadas (por ejemplo, 2.4E+16). Este problema se debía a una incompatibilidad entre los formatos de separador decimal utilizados por Python (punto ".") y Power BI (coma ","), que provocaba que valores como 24.83 fueran interpretados como 2.483.000.000.000.000.

La solución fue recalcular la columna dentro de Power BI mediante el editor de Power Query, usando las variables originales cases y py con la fórmula nombrada anteriormente.

Este campo se renombró como incidence_rate_new y fue utilizado en todas las visualizaciones posteriores.

El dashboard se estructuró en tres páginas. La primera sobre las tendencias temporales y por edad, donde se representó la evolución de la tasa de incidencia del cáncer cerebral desde 1953 hasta 2017, diferenciando por sexo. Se observó una ligera diferencia constante a lo largo del tiempo, siendo los hombres quienes presentan mayores tasas. Además, se visualizó la evolución por grupos de edad, mostrando un claro aumento progresivo de la incidencia en edades avanzadas, especialmente a partir de los 65 años.

La segunda sobre la variabilidad según sexo y subtipo tumoral, donde se muestra que los hombres presentan una tasa media superior (0,66 frente a 0,50 en mujeres), manteniéndose esta diferencia en todos los análisis. También se construyó un gráfico de sectores con los subtipos histológicos de cáncer cerebral, donde destacan los tumores astrocíticos (32,7%), seguidos de morfología no especificada y gliomas de origen incierto.

Y, por último, distribución geográfica por continente donde se analizó la tasa media de incidencia por continente, destacando Europa como la región con mayor incidencia,

seguida por Oceanía y Sudamérica. Esta variabilidad geográfica puede explicarse por factores ambientales, genéticos o por diferencias en la calidad y cobertura de los registros sanitarios.

11. Visualización interactiva con Streamlit

Además del análisis y las visualizaciones desarrolladas en Power BI, se construyó una aplicación interactiva en Streamlit con el objetivo de explorar los datos de manera más dinámica y flexible. Esta herramienta permitió crear un entorno accesible desde cualquier navegador.

A través de esta aplicación se incluyeron visualizaciones clave como un histograma por grupos de edad para observar cómo se distribuyen los casos, un mapa de calor que muestra las correlaciones entre variables como la edad, el número de casos, los años-persona y la tasa de incidencia, y varios gráficos de barras que permiten comparar la carga de la enfermedad por sexo y por tipo de tumor. También se diseñaron visualizaciones interactivas como un gráfico de líneas que analiza la evolución de la incidencia según la edad y el sexo, y un diagrama de dispersión que permite explorar la relación entre edad, incidencia, sexo y subtipo histológico. Para completar la parte geográfica, se incorporó un mapa con las tasas medias de incidencia por continente, junto con un panel comparativo que resume los principales indicadores por región.

12. Conclusión

Este proyecto ha permitido explorar en profundidad la incidencia del cáncer cerebral a nivel global a través del análisis de datos reales, aplicando técnicas de procesamiento, visualización y modelado estadístico. Gracias al uso combinado de Python, Power BI y Streamlit, fue posible no solo identificar patrones relevantes según edad, sexo, subtipo tumoral y continente, sino también comunicar los resultados de manera clara y accesible.

Los hallazgos reflejan importantes diferencias epidemiológicas: la incidencia tiende a aumentar con la edad, es ligeramente más alta en hombres y presenta variaciones significativas según el tipo de tumor y la región geográfica. Estas desigualdades podrían estar vinculadas a factores biológicos, ambientales o incluso a diferencias en los sistemas de registro y diagnóstico.

Más allá de los resultados específicos, el valor principal de este trabajo radica en su enfoque integrador. La combinación de ciencia de datos con herramientas visuales interactivas demuestra cómo el análisis de grandes volúmenes de información puede convertirse en una herramienta potente para generar conocimiento, apoyar la toma de decisiones en salud pública y abrir nuevas líneas de investigación.