

## Ajustes de análisis – bitácora con correcciones

### 1. Tema analítico (Pregunta de investigación):

¿Existe asociación entre las concentraciones mensuales de PM2.5 y la mortalidad prematura por enfermedad cardiocerebrovascular (30–70 años) en las localidades de Bogotá D.C. durante el año 2021, y en qué medida difiere dicha asociación?

### 2. Análisis requeridos o inferidos (selección de variables):

#### Variables principales.

- **Variable dependiente:** Número de muertes prematuras por enfermedad cardiocerebrovascular. Esta variable ha sido validada para ser usada en la investigación. Se cuenta con registros mensuales para varios años de muertes de personas causadas por el tipo de enfermedades mencionado, por lo que se puede hacer el respectivo conteo de las muertes.

- **Variable independiente:** PM2.5. La variable independiente principal ha sido validada para ser usada. Se cuentan con los registros de esta variable proporcionados por los datos base con los que se cuenta para el proyecto. Esta es la principal variable explicativa para la investigación con la que se pretende mostrar posibles efectos de la contaminación ambiental sobre las muertes prematuras por enfermedades cardiocerebrovasculares.

#### Otras variables.

- **Temperatura promedio de Bogotá:** Se ha validado esta variable para ser usada en el proyecto. Está disponible un dataset con datos de la temperatura promedio de Bogotá mensual para varios años desde el 2007, lo cual favorece la temporada que se pretende analizar en el proyecto. Con esto en mente, la variable temperatura resulta relevante en la investigación para considerar otras condiciones ambientales que tengan algún efecto sobre las muertes y que, además, pueden afectar la asociación entre las variables principales del proyecto.

- **Sexo:** En el mismo dataset en el que se encuentra la información de la variable dependiente, se encuentra la información del sexo de las personas para las que se registró su muerte por enfermedades cardiocerebrovasculares. Por ende, se valida que se puede usar esta variable dentro de la investigación. Además, la variable resulta relevante para considerar en la investigación si la biología humana tiene alguna afectación sobre la asociación principal.

- **Edad:** Nuevamente, se cuenta con la información de la edad de las personas para las cuales se registró la muerte. Por tal motivo, se valida que la variable se puede usar para la construcción del modelo para proponer en la investigación. En ese sentido, la variable puede proporcionar información valiosa sobre si la edad de una persona puede afectar el efecto de la contaminación ambiental sobre las muertes prematuras por el tipo enfermedades de interés.

- **Localidad:** Se valida el uso de esta variable para el proyecto de investigación porque se cuenta con la información de esta en el dataset de registros de las personas que murieron por enfermedades cardiocerebrovasculares. Además, la variable es valiosa para discriminar la asociación de interés del proyecto de acuerdo con condiciones sobre las localidades como, por ejemplo, una mayor concentración vehicular en localidades más grandes como Kennedy.

Adicionalmente, los requerimientos analíticos se complementan con los siguientes ítems para lograr con el proyecto.

- Determinar la incidencia de la concentración de material particulado PM2.5 en el aire sobre las muertes prematuras por enfermedades cardiocerebrovasculares.
- Comparar el comportamiento de la variable de interés del proyecto con base en las distintas localidades de Bogotá en las que ocurren las muertes que se pretenden analizar.
- Discriminar las muertes prematuras por enfermedades cardiocerebrovasculares de acuerdo con el sexo y la edad de la persona afectada.
- Observar diferencias en el comportamiento de la variable dependiente de interés a través del tiempo.
- Evidenciar el efecto de las condiciones ambientales, particularmente la temperatura, sobre las muertes prematuras por las enfermedades cardiocerebrovasculares.

### **3. Justificación desde el punto de vista clínico:**

Se decidió estudiar esta asociación porque las enfermedades cardiocerebrovasculares representan una de las principales causas de mortalidad prematura en Colombia y el mundo, y se ha identificado que la contaminación atmosférica, en particular la exposición a partículas finas PM<sub>2.5</sub>, actúa como un determinante ambiental clave que agrava este tipo de enfermedades. En ciudades como Bogotá, donde existen diferencias marcadas en las condiciones ambientales entre localidades, estudiar esta relación permite comprender mejor cómo la exposición a contaminantes del aire contribuye a las desigualdades en salud.

El tema se aborda mediante un análisis ecológico comparativo entre las localidades de Fontibón y Kennedy, que aunque están geográficamente próximas, muestran comportamientos contrastantes en mortalidad prematura y niveles de contaminación. Al evaluar la asociación entre las concentraciones mensuales de PM<sub>2.5</sub> y la mortalidad prematura por enfermedad cardiocerebrovascular (30–70 años) durante el año 2021, se busca evidenciar si las diferencias en mortalidad pueden explicarse, al menos en parte, por variaciones en la exposición ambiental.

Desde el punto de vista clínico, numerosos estudios han demostrado que la exposición prolongada a PM<sub>2.5</sub> se asocia con un mayor riesgo de enfermedad isquémica del corazón, accidente cerebrovascular y mortalidad cardiovascular. El metaanálisis de Alexeeff et al. (2021), publicado en Journal of the American Heart Association, encontró que un incremento de 10 µg/m<sup>3</sup> en PM<sub>2.5</sub>

se asocia con un aumento del 23% en la mortalidad por enfermedad isquémica y del 24% en la mortalidad cerebrovascular, confirmando su papel como factor de riesgo modificable.

Asimismo, las revisiones sistemáticas de Gong et al. (2018) y Wang et al. (2023) respaldan esta evidencia, mostrando que tanto la exposición crónica como la exposición a corto plazo a PM<sub>2.5</sub> se relacionan con un incremento significativo en los eventos cardiovasculares, especialmente en zonas urbanas con altos niveles de contaminación.

Por tanto, este trabajo aborda un problema de relevancia clínica y de salud pública, al intentar relacionar un contaminante atmosférico de importancia global con la mortalidad prematura local, aportando evidencia útil para la gestión ambiental y la prevención de enfermedades crónicas no transmisibles.

### *Referencias*

1. Alexeeff, S. E., Roy, A., Shan, J., Liu, X., Messier, K. P., Thelen, J., Apte, J. S., & Karagueuzian, K. G. (2021). Long-Term PM<sub>2.5</sub> Exposure and Risks of Ischemic Heart Disease and Stroke Events: Review and Meta-Analysis. *Journal of the American Heart Association*, 10(9): e019751.
2. Gong, T., et al. (2018). Long-term exposure to ambient PM<sub>2.5</sub> and cardiovascular disease: A systematic review and meta-analysis. *Environmental Research*, 167, 448–460.
3. Wang, M., et al. (2023). Long-term exposure to ambient PM<sub>2.5</sub> and the incidence of cardiovascular and cerebrovascular diseases: An updated systematic review and meta-analysis. *Atherosclerosis*, 375, 67–79.

## **4. Justificación desde el punto de vista técnico**

La evaluación de la asociación entre la concentración de material particulado fino (PM<sub>2.5</sub>) y la mortalidad prematura por enfermedad cardiocerebrovascular brinda un escenario adecuado para el uso de modelos analíticos robustos para estimar los efectos a corto plazo de la contaminación atmosférica sobre la mortalidad. Con esto en mente, un modelo que puede resultar pertinente para el caso es el modelo de rezago no lineal distribuido (DLNM). Este tipo de modelos facilita la captura de la naturaleza retardada y potencialmente no lineal de la asociación para investigar. Adicionalmente, la integración de la analítica de datos con el ámbito clínico favorece la interpretación de los datos poblacionales sobre la salud de las personas, así como, la toma de decisiones acertadas en la salud pública. De este modo, la construcción de herramientas de inteligencia de negocios para este tipo de investigaciones contribuye al diseño de intervenciones basadas en evidencia por parte de los expertos del campo (Gasparrini et al., 2010).

### *Referencias*

- Gasparrini, A., Armstrong, B., & Kenward, M. G. (2010). Distributed lag non-linear models. *Statistics in Medicine*, 29(21), 2224–2234. <https://doi.org/10.1002/sim.3940>

## 5. Objetivo principal y secundarios

### Objetivo general:

Evaluar la asociación entre las concentraciones mensuales de PM2.5 y la mortalidad prematura por enfermedad cardiocerebrovascular (30–70 años) en las localidades de Fontibón y Kennedy, Bogotá D.C., durante el año 2021.

### Objetivos específicos:

- Cuantificar la asociación entre las concentraciones mensuales de PM2.5 y el número de muertes prematuras por enfermedad cardiocerebrovascular en cada localidad (Fontibón y Kennedy), ajustando por temperatura, edad, sexo y localidad.
- Comparar la magnitud y significancia de la asociación PM2.5 y la mortalidad entre Fontibón y Kennedy mediante análisis estratificado o de interacción por localidad.
- Analizar la variación temporal mensual de las concentraciones de PM2.5 y su relación con los cambios en la mortalidad prematura durante el año 2021.

## 6. Fuentes de datos complementarios

Para el desarrollo del proyecto se utilizan fuentes de datos abiertas y verificables que permiten integrar variables ambientales, demográficas y de salud pública con el fin de evaluar la asociación entre la contaminación atmosférica y la mortalidad prematura. En este sentido, los conjuntos de datos seleccionados complementan la información principal de manera metodológicamente sólida y relevante para el contexto de Bogotá D.C.

Mortalidad prematura por enfermedad cardiocerebrovascular en Bogotá D.C. (30 a 70 años):

<https://datosabiertos.bogota.gov.co/dataset/mortalidad-prematura-por-enfermedad-cardiocerebrovascular-en-bogota>

Temperatura en Bogotá D.C. y su relación con ENOS:

<https://datosabiertos.bogota.gov.co/dataset/temperaturas-en-bogota-d-c>

## Análisis para el tablero de control inferencial

### Análisis estadístico:

El tablero de control inferencial propuesto tiene como objetivo evaluar la posible asociación entre las concentraciones mensuales de PM2.5 y la mortalidad prematura por enfermedad cardiocerebrovascular en personas de 30 a 70 años residentes en las diferentes localidades de Bogotá D.C. durante el año 2021. Este tablero permitirá realizar análisis estadísticos dinámicos que se ajusten automáticamente a los filtros y parámetros seleccionados por el usuario, ofreciendo resultados claros y comparables según distintas condiciones de exposición y contexto sociodemográfico.

El diseño del tablero se basará en un sistema interactivo que posibilite seleccionar el nivel de confianza (entre el 90% y el 99%) y el nivel de significancia ( $\alpha$ ) (entre 0.01 y 0.1). Estas opciones garantizarán que el usuario pueda adaptar el grado de precisión y exigencia estadística de los resultados según sus necesidades. Los filtros principales incluirán localidad, mes, sexo, rango de edad, nivel de exposición a PM2.5 y variables contextuales como el índice IBOCA, que mide la calidad ambiental territorial.

### **Medida estadística principal:**

El análisis inferencial se centrará en la estimación del Hazard Ratio (HR) mediante un modelo de regresión de Cox, dado que la variable dependiente corresponde a la ocurrencia del evento (muerte prematura) y al tiempo hasta su aparición. Este enfoque de análisis de supervivencia resulta el más adecuado para este tipo de investigación, ya que permite estimar el riesgo instantáneo de mortalidad en función de la exposición a PM2.5, controlando simultáneamente por otras variables relevantes.

El HR ofrece una medida robusta y fácilmente interpretable del efecto de la contaminación ambiental sobre la mortalidad prematura, al comparar la probabilidad de morir entre grupos con diferentes niveles de exposición. El modelo ajustará además por edad, sexo, localidad e índice IBOCA, con el fin de eliminar el efecto de posibles factores de confusión.

### **Estructura de resultados en el tablero:**

Cada vez que el usuario aplique filtros o modifique parámetros, el tablero recalculará automáticamente los resultados del análisis. Para cada subconjunto de datos seleccionado, se mostrará:

- El valor estimado del HR asociado a la exposición a PM2.5.
- Su intervalo de confianza, ajustado al nivel elegido (90%, 95% o 99%).
- El valor p de la prueba de hipótesis correspondiente, comparado con el nivel de significancia definido por el usuario.
- El tamaño de la muestra, que incluirá tanto el número total de observaciones (individuos o unidades localidad–mes) como el número de eventos (defunciones).

De esta manera, el tablero permitirá evaluar no solo la significancia estadística de la asociación, sino también la solidez de las estimaciones, reflejada en el número de datos disponibles y en la amplitud de los intervalos de confianza. Cuando el número de eventos sea insuficiente para obtener una estimación confiable, el sistema emitirá un mensaje de advertencia, recomendando interpretar los resultados con precaución o reagrupar categorías.

### **Análisis complementarios:**

Además del modelo principal basado en el HR, el tablero incluirá herramientas para calcular y comparar promedios y proporciones, con sus respectivos intervalos de confianza, de acuerdo con los filtros seleccionados. Por ejemplo, se podrá comparar la media mensual de PM2.5 entre

diferentes localidades o entre grupos de exposición (bajo, medio y alto), así como la proporción de muertes observadas en cada grupo.

Para este tipo de comparaciones, el tablero aplicará pruebas de hipótesis adecuadas: la prueba t de Student en el caso de diferencias de medias (o su alternativa no paramétrica cuando no se cumplan los supuestos de normalidad) y la prueba de chi-cuadrado o Fisher exacta en el caso de proporciones. En ambos casos, se calculará el intervalo de confianza de la diferencia de medias o proporciones al nivel de confianza definido por el usuario, y se mostrará el valor p obtenido junto con la decisión estadística (rechazo o no rechazo de la hipótesis nula) según el nivel de significancia seleccionado.

Estas pruebas complementarias permitirán contextualizar los hallazgos del modelo de Cox, aportando evidencia adicional sobre la existencia de diferencias significativas entre grupos de exposición o localidades. En conjunto, las comparaciones descriptivas y los modelos de supervivencia brindarán una visión más completa de cómo varía el riesgo de mortalidad en función de la contaminación ambiental.

### **Análisis de los coeficientes del modelo de regresión:**

El tablero incluirá una sección que presentará los coeficientes obtenidos del modelo de regresión multivariado, permitiendo examinar la magnitud y dirección del efecto de cada variable sobre la mortalidad prematura. Allí se mostrarán, para cada variable incluida, el coeficiente estimado, el HR correspondiente, el intervalo de confianza según el nivel seleccionado y el valor p asociado a su prueba de hipótesis. También se incluirán indicadores globales de ajuste del modelo, así como medidas de correlación entre variables (por ejemplo, índice de colinealidad), con el fin de garantizar la validez y solidez del análisis.

El modelo de Cox permitirá, además, evaluar el aporte individual de variables como el nivel de PM2.5, el IBOCA, la edad, el sexo y la localidad, así como su interacción. Esto permitirá identificar si el efecto de la contaminación difiere entre subgrupos (por ejemplo, entre hombres y mujeres, o entre localidades con distintas condiciones ambientales). Para visualizar estas diferencias, se emplearán gráficos tipo *forest plot*, que mostrarán los HR y sus intervalos de confianza para cada categoría, facilitando la comparación visual de los resultados.

### **Pruebas de hipótesis e interpretación:**

El tablero evaluará tres tipos principales de hipótesis. En primer lugar, la hipótesis nula que plantea que no existe asociación significativa entre la exposición a PM2.5 y la mortalidad prematura por enfermedad cardiocerebrovascular, frente a la hipótesis alternativa, que sostiene que sí existe una asociación estadísticamente significativa. Esta evaluación se realizará comparando el valor p del coeficiente de exposición con el nivel de significancia elegido. Si p es menor que  $\alpha$ , se rechazará la hipótesis nula, concluyendo que la exposición a PM2.5 tiene un efecto significativo sobre la mortalidad prematura.

En segundo lugar, se aplicarán pruebas para comparar los promedios o porcentajes de mortalidad entre distintos grupos, como localidades o niveles de exposición. Si los resultados muestran

diferencias significativas al nivel de confianza seleccionado, se interpretará que las tasas de mortalidad difieren según el grado de exposición al contaminante. Finalmente, el tablero permitirá evaluar la homogeneidad de los efectos entre subgrupos mediante pruebas de interacción, determinando si la magnitud del HR cambia entre categorías. Si los intervalos de confianza de los HR se solapan ampliamente o la prueba de interacción no es significativa, se interpretará que el efecto del PM2.5 es estable entre los subgrupos analizados.

### **Visualización y salida de resultados:**

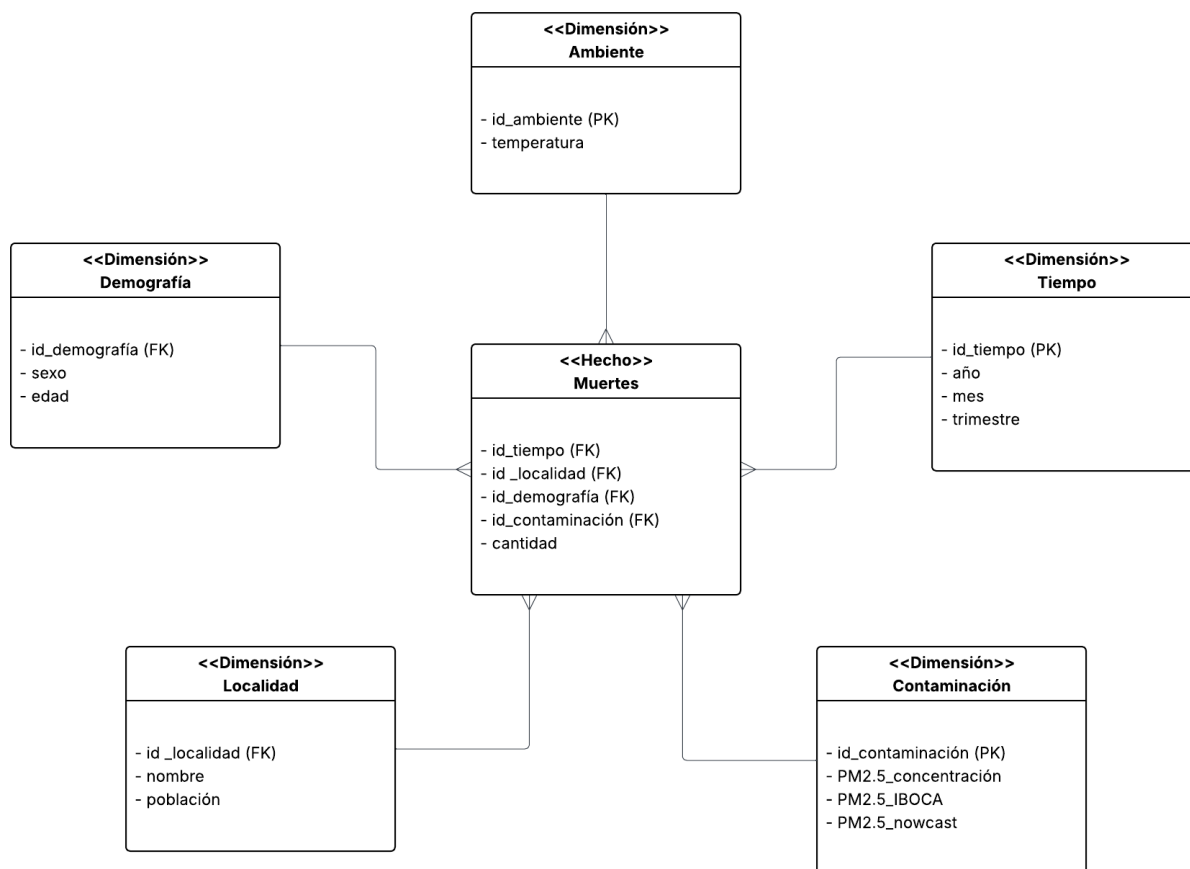
El tablero presentará los resultados de forma visual y comprensible para facilitar su interpretación. Cada análisis mostrará tablas con los valores de HR, sus intervalos de confianza y los valores p correspondientes, acompañadas del tamaño muestral total y del número de eventos. Además, se incluirán gráficos de tendencia temporal entre las concentraciones de PM2.5 y las defunciones mensuales, mapas temáticos que muestren las diferencias espaciales entre localidades y *forest plots* que permitan comparar los HR entre distintos grupos.

Cada salida estará acompañada de una interpretación textual automática. Por ejemplo, si se obtiene un HR de 1.15 con un intervalo de confianza al 95% de 1.06 a 1.25 y un valor p de 0.004, el tablero mostrará el mensaje: “Existe evidencia estadística de que mayores niveles de PM2.5 se asocian con un mayor riesgo de mortalidad prematura por enfermedad cardiocerebrovascular ( $\alpha=0.01$ ).” Si, por el contrario, el HR es 1.04 con un intervalo que incluye la unidad y un valor p superior al nivel de significancia seleccionado, el tablero mostrará: “No se encontró evidencia estadísticamente significativa de asociación entre PM2.5 y mortalidad prematura en las condiciones analizadas.”

El tablero de control inferencial combinará herramientas estadísticas avanzadas con una interfaz flexible e interactiva que permitirá analizar la relación entre contaminación ambiental y mortalidad de manera integral. Al incorporar modelos de supervivencia (basados en HR), comparaciones de medias y proporciones, análisis por subgrupos y representaciones gráficas, el tablero ofrecerá una visión completa y dinámica del comportamiento de la mortalidad prematura frente a las concentraciones de PM2.5 en Bogotá durante 2021.

La posibilidad de ajustar el nivel de confianza y de significancia, junto con la visualización del tamaño muestral y los intervalos de confianza, garantizará la transparencia y precisión de los resultados. Además, la inclusión de diagnósticos de modelo y advertencias sobre limitaciones estadísticas permitirá mantener la validez de las conclusiones. En última instancia, este tablero no solo proporcionará evidencia estadística rigurosa sobre la asociación entre contaminación y mortalidad, sino que también servirá como una herramienta práctica para la toma de decisiones en salud pública y planificación ambiental en la ciudad.

## Modelo multidimensional



El modelo multidimensional propuesto para el proyecto de investigación se basa en el modelado de un único hecho, las muertes. Este hecho corresponde con el principal proceso dentro del contexto de la pregunta de investigación abordada que se quiere explicar. De este modo, se plantea como único hecho de interés en el modelo las muertes prematuras causadas por enfermedades cerebrocardiovasculares en Bogotá.

En este orden de ideas, el modelo multidimensional incluye una serie de dimensiones que tienen relaciones con el hecho. La inclusión de dichas dimensiones tiene el propósito de aportar información relevante, de aportar contexto, a la ocurrencia del hecho de las muertes. De esta manera, se busca modelar el aporte de información por parte de las dimensiones, de tal forma que, se pueda explicar la cantidad de muertes mediante los atributos de las dimensiones y, así, dar una respuesta efectiva a la pregunta de investigación.

Con esto en mente, las dimensiones incluidas en el modelo son 5. La primera dimensión es la dimensión del tiempo. Mediante esta dimensión se pretende obtener información sobre el número de muertes prematuras causadas por enfermedades cardiocerebrovasculares a través del tiempo. La segunda dimensión corresponde a la de la localidad (geografía). El objetivo de la inclusión de esta dimensión es poder analizar las muertes ocurridas en las distintas localidades de Bogotá, es decir, analizar cómo cambia el comportamiento de las muertes de la pregunta de investigación de acuerdo con una discriminación geográfica de la ciudad mediante las localidades que forman Bogotá. La tercera dimensión es la dimensión de la contaminación. Con dicha dimensión se busca explicar los acontecimientos de interés mediante la principal variable explicativa planteada en la pregunta de investigación, que corresponde a la concentración de material particulado PM2.5. La cuarta dimensión corresponde a la dimensión de la demografía. A través de la dimensión demográfica, se pretende aportar una mayor explicación a las muertes prematuras causadas por las enfermedades de interés con base en las características de las personas afectadas, principalmente, su edad y su sexo. Por último, se incluye la dimensión del ambiente. Con la inclusión de esta dimensión, se pretende aportar información valiosa sobre las condiciones ambientales al momento de la ocurrencia de los acontecimientos de interés, es decir, las muertes. Principalmente, se pretende aportar información sobre la temperatura.

Por último, se listan los atributos incluidos en el hecho y en cada una de las dimensiones que forman el modelo propuesto. Asimismo, se da una breve explicación de cada atributo con el fin de dar mayor claridad en el entendimiento del modelo.

#### - Muertes (hecho):

- **id\_tiempo:** Llave foránea que permite establecer la relación 1-n por parte de la dimensión del tiempo con el hecho de las muertes.
- **id\_localidad:** Llave foránea que permite establecer la relación 1-n por parte de la dimensión de la localidad con el hecho de las muertes.
- **id\_demografía:** Llave foránea que permite establecer la relación 1-n por parte de la dimensión de la demografía con el hecho de las muertes.
- **id\_contaminación:** Llave foránea que permite establecer la relación 1-n por parte de la dimensión de la contaminación con el hecho de las muertes.
- **id\_ambiente:** Llave foránea que permite establecer la relación 1-n por parte de la dimensión del ambiente con el hecho de las muertes.
- **cantidad:** Atributo que indica la cantidad de muertes prematuras que ocurren por enfermedades cardiocerebrovasculares. De acuerdo con la pregunta de investigación, el registro es mensual.

#### - Tiempo (dimensión):

- **id\_tiempo:** Llave primaria subrogada que permite identificar un registro de la dimensión.
- **año:** Año en el que ocurrió el acontecimiento de interés.
- **mes:** Mes en el que ocurrió el acontecimiento de interés.

- **trimestre:** Trimestre en el que ocurrió el acontecimiento de interés. Se incluye este atributo con el fin de obtener una mayor visión de análisis que no se limite a la temporalidad mensual.

**- Localidad (dimensión):**

- **id\_localidad:** Llave primaria subrogada que permite identificar un registro de la dimensión.
- **nombre:** Nombre de la localidad en la que ocurrió el evento de interés.
- **población:** Población de la localidad en la que ocurrió el evento de interés.

**- Contaminación (dimensión):**

- **id\_contaminación:** Llave primaria subrogada que permite identificar un registro de la dimensión.
- **PM2.5\_concentración:** Concentración del agente contaminante en  $\mu\text{g}/\text{m}^3$  (microgramos por metro cúbico).
- **PM2.5\_IBOCA:** Registro de la calidad del aire con el sistema IBOCA.
- **PM2.5\_nowcast:** Registro de la calidad del aire con el método nowcast.

**- Demografía (dimensión):**

- **id\_demografía:** Llave primaria subrogada que permite identificar un registro de la dimensión.
- **sexo:** Sexo de la persona que murió prematuramente por enfermedades cerebrocardiovasculares.
- **edad:** Edad, al momento de morir, de la persona que murió prematuramente por enfermedades cerebrocardiovasculares.

**- Ambiente (dimensión):**

- **id\_ambiente:** Llave primaria subrogada que permite identificar un registro de la dimensión.
- **temperatura:** Temperatura promedio registrada por Bogotá en el momento de la ocurrencia de los acontecimientos (muertes) de interés.