

**UNIVERSIDAD NACIONAL DE ENTRE RÍOS FACULTAD DE INGENIERÍA**

**Tecnicatura Universitaria en Procesamiento y Explotación de Datos**

**Bases de Datos Multidimensionales**

**Suicidios en Argentina 2017-2022: Descifrando las Sombras**

**Colignon, Sabrina**

**2 de Noviembre de 2023**

# 

# Índice

[**Índice 1**](#_q3lv33j5uit4)

[**Introducción 2**](#_byk9p5cuzqnl)

[**Situación problemática 2**](#_v8oq0d9talt3)

[Planteo de la situación problemática 2](#_hj04ezrkv9an)

[**Marco teórico 3**](#_2zpe2x7eyzkl)

[Estado del arte 3](#_34r3cocc0ioz)

[Bases teóricas 4](#_k9v18roko4vt)

[**Diseño e implementación 6**](#_lc36rmjxn01u)

[ETL 9](#_cibgv1jkmtm9)

[Operaciones OLAP 11](#_4cetb5xtm5rv)

[Plan de mantenimiento 12](#_op1ylk3i0nsk)

[Reportes 12](#_ebledel03a7n)

[**Conclusión 14**](#_2pa5gsu28qh5)

[**Anexo 16**](#_yznls6qu2j5w)

[Diccionario de datos 16](#_hyc5dz9w6wph)

# 

# Introducción

Este informe presenta información útil para la posible formulación y gestión de políticas públicas, como resultado de la identificación de patrones y tendencias de suicidas a lo largo de todo el territorio nacional en los años comprendidos entre 2017 y 2022, teniendo en cuenta su edad, sexo, lugar, fecha y modalidad. Como se mostrará a lo largo del documento, se trabajó a partir de la base de datos sobre “Hechos y víctimas de suicidios en la República Argentina provenientes del Sistema de Alerta Temprana del Sistema Nacional de Información Criminal desde el año 2017 hasta el año 2022”, se utilizó PgAdmin 4 para generar las tablas de la base de datos multidimensional, Visual Studio Code para la codificación del ETL y PowerBi para la realización del dashboard para visualizar la información.

# Situación problemática

El suicidio es el evento mediante el cual una persona de manera deliberada se quita la vida. Según la Organización Mundial de la Salud (OMS) cada año, cerca de 703.000 personas se quitan la vida y muchas más intentan hacerlo. Todos los casos son una tragedia que afecta a familias, comunidades y países y tienen efectos duraderos para los allegados de la víctima. Puede ocurrir a cualquier edad, y en 2019 fue la cuarta causa de defunción en el grupo etario de 15 a 29 años en todo el mundo.

Los suicidios no solo ocurren en los países de altos ingresos, sino que es un fenómeno que afecta a todas las regiones del mundo. De hecho, más del 77% de los suicidios ocurridos en 2019 tuvieron lugar en países de ingresos bajos y medianos.

Es por esto que el Ministerio de Seguridad junto con el Ministerio de Salud de la Nación Argentina estan buscando realizar una nueva campaña sobre prevencion del suicidio y para eso cuentan con una base de datos sobre hechos y víctimas de suicidios en la República Argentina provenientes del Sistema de Alerta Temprana del Sistema Nacional de Información Criminal desde el año 2017 hasta el año 2022. A partir de estos datos buscan responder una serie de preguntas para concientizar y hablar sobre esta problemática a nivel nacional.

## Planteo de la situación problemática

El dataset provisto por el Ministerio de Seguridad cuenta con los datos provenientes de suicidios registrados a lo largo del país, cada uno cuenta con un identificador, fecha, hora, lugar, modalidad, motivo, edad y sexo. Asimismo las variables están asociadas a un código por departamento y localidad dónde ocurrió el hecho. Corresponde a los hechos consumados entre los años 2017 y 2022 inclusive, contiene 21720 registros y 23 filas. Se relevan únicamente aquellos hechos consumados.

“La base de datos cuenta con un identificador del hecho, para cada hecho se encuentran en la misma base los registros correspondientes a las personas involucradas. En esta base usuaria sólo se incluye información sobre las víctimas (no sobre testigos); en caso de hechos con más de una víctima los datos del hecho se repiten para cada víctima. Para conocer la cantidad de hechos, se debe contar una única vez cada ID de hecho” Manual de usuario Base SAT-SS (2023), Ministerio de Seguridad de la Nación. Dirección Nacional de Estadística Criminal.

Los datos se encuentran en un archivo de tipo CSV (comma-separated values) que fue descargado del Portal Nacional de Datos Públicos (<https://www.datos.gob.ar/>) que dispone datos públicos en formatos abiertos para que puedan ser usados, modificados y compartidos. El diccionario de datos extraído del sitio está disponible en el Anexo 1.

A partir del análisis del dataset los indicadores más importantes para resolver las interrogantes son edad, sexo, lugar, fecha y modalidad. Teniendo estos indicadores en cuenta se busca responder a la siguiente interrogante:

¿Cuál sería la metodología adecuada para examinar los patrones y tendencias de la base de datos de suicidios a nivel nacional, con el propósito de informar el diseño de una nueva campaña de prevención del suicidio que será implementada por el Ministerio de Salud y el Ministerio de Seguridad en el período comprendido entre 2017 y 2022?

# Marco teórico

## Estado del arte

El suicidio constituye una de las principales causas de mortalidad en diversos países, sin hacer distinción entre naciones ricas y pobres.

La fiabilidad de los datos es cuestionable debido a la estigmatización asociada a este fenómeno, lo que conduce a ocultar, en muchos casos, la verdadera causa del fallecimiento, y no se registra como suicidio.

La Organización Mundial de la Salud (OMS) sostiene que es posible prevenir los suicidios mediante medidas que restrinjan el acceso a los métodos más comunes de autoeliminación, políticas destinadas a reducir el consumo de alcohol y la atención y tratamiento de personas con problemas de salud mental, enfermedades crónicas y adicciones.

En el grupo de jóvenes de 15 a 29 años, el suicidio ocupa el cuarto lugar como causa principal de muerte, después de los traumatismos causados por accidentes de tráfico, la tuberculosis y la violencia interpersonal. Las tasas varían entre países, regiones y géneros.

Según la OMS el número de suicidios masculinos duplica al de las mujeres (12,6 por cada 100.000 hombres frente a 5,4 por cada 100.000 mujeres). Las tasas de suicidio masculino suelen ser más altas en países de ingresos elevados (16,5 por cada 100.000), mientras que las tasas más altas de suicidio femenino se encuentran en países de ingresos medios y bajos (7,1 por cada 100.000).

A nivel mundial en 2019, las tasas de suicidio en las regiones de África (11,2 por cada 100.000), Europa (10,5 por cada 100.000) y Asia Sudoriental (10,2 por cada 100.000) de la OMS superaron la media global (9,0 por cada 100.000), con la tasa de suicidio más baja registrada en la Región del Mediterráneo Oriental (6,4 por cada 100.000).

A pesar de que algunos países han priorizado la prevención del suicidio en sus programas, son muchos los que no han adoptado medidas al respecto. Actualmente, solo 38 países cuentan con una estrategia nacional de prevención del suicidio. Es esencial promover considerablemente las medidas en este ámbito para cumplir con el objetivo de reducir en un tercio la tasa global de suicidios para 2030, como se establece en los Objetivos de Desarrollo Sostenible (ODS).

En Argentina, se registra un fallecimiento por suicidio cada tres horas, de acuerdo al informe más reciente de la Dirección de Estadísticas e Información de Salud, una entidad bajo el Ministerio de Salud de la Nación. Durante el año 2021, se documentaron un total de 2,865 casos de suicidio, lo que representa una tasa de 6 suicidios por cada 100,000 habitantes.

A pesar de ser un serio desafío en el ámbito de la salud pública, es posible prevenirlo mediante medidas efectivas respaldadas por pruebas y, en muchas ocasiones, de bajo costo. Para que las iniciativas de prevención a nivel nacional sean eficaces, es esencial adoptar un enfoque integral y multisectorial.

## Bases teóricas

Un Data Warehouse es un depósito electrónico en el cual una empresa u organización suele almacenar una extensa cantidad de información centrada en un tema específico. En su definición clásica, un Data Warehouse se caracteriza por:

* Orientado a temas: Los datos de una base de datos se organizan de manera que todos los elementos relacionados con el mismo evento u objeto del mundo real están vinculados entre sí, de modo que cada característica devuelva datos coherentes.
* Variabilidad en el tiempo: Los cambios en los datos a lo largo del tiempo se registran para permitir la generación de informes que reflejen estas variaciones.
* No volatilidad: La información almacenada no se modifica ni se elimina; una vez que un dato se almacena, se convierte en información de solo lectura y se mantiene para futuras consultas.
* Integración: La base de datos contiene datos de todos los sistemas operativos de la organización, y se asegura que estos datos sean coherentes entre sí.

Comenzando con el diseño de nuestro datawarehouse, tenemos tres tipos. El primero es el diseño conceptual, en bases de datos multidimensionales no hay una convención por ende pueden ser “copo de nieve” o “estrella”. Asimismo el nivel conceptual se divide en el nivel conceptual y en el nivel conceptual ampliado. En el primer caso ya no se trabaja en términos de entidades y relaciones sino en términos de hechos y dimensiones, se definen las dimensiones mediante las cuales se agruparán las medidas. luego se define el hecho y por último se establecen las medidas que definen los hechos. Si existieran más hechos, tendríamos una constelación de hechos. En el caso del diseño conceptual ampliado, se define el origen desde donde se obtienen las dimensiones y se define el origen y la función de agregación de las medidas.

Ahora bien un esquema está compuesto por dimensiones y en el centro tiene un cubo (donde cada cara del cubo corresponde a una dimensión). Una dimensión está compuesta por niveles, define como están los datos organizados lógicamente y proveen el medio para analizar el contexto de negocios. Podemos tener datos cualitativos, aspectos de interés.

Para que una dimensión sea exitosa:

* Utilizar textos completos entendibles
* Muy descriptiva
* Completa (sin omitir valores)
* Asegurar calidad (bien escrito, sin valores imposibles, sin valores obsoletos, sin diferentes versiones del mismo atributo)
* Indexación
* Desnormalizada: para encontrar las dimensiones y niveles.
* Bien documentada

Los niveles son análogos a las entidades de un der o similares. Podemos tener representados conceptos “granulados”. podemos tener datos cuantitativos. Por último los hechos son instancias que se hacen de un nivel y esos hechos se usan para formar los indicadores. Los indicadores tienen métricas. Mientras mayor sea la cantidad de dimensiones, más específica es la información.

En cuanto a la jerarquía son elementos clave en las aplicaciones analíticas, ya que proporcionan los medios para representar los datos bajo análisis en diferentes niveles de abstracción. Pueden ser

1. Equilibrada: solo una ruta de esquema, donde todos los niveles son obligatorios, tengo que pasar por cada instancia para subir a un nivel
2. No equilibrada: no todos los niveles son obligatorios.
3. Paralela (Independiente - Dependiente)
4. Generalizada (Sin nivel de unión - Desiguales/Desconectadas)
5. Estricta
6. No estricta
7. Alternativa

Continuando con el diseño lógico, en la tabla de hechos aparecerán las foreign keys a cada tabla de dimensiones y los atributos que representan las medidas. En esta etapa del diseño se decide si se utilizarán jerarquías copo de nieve o estrella. En caso que existan múltiples hechos (constelación) se definen las tablas, las cuales compartirán dimensiones.

El diseño lógico en bases de datos multidimensionales se centra en la estructura y organización de la base de datos en términos de dimensiones, jerarquías, hechos, cubos y relaciones entre los datos. Retomando, las dimensiones representan categorías clave, como tiempo, ubicación, producto, persona, entre otras. Las jerarquías describen las relaciones entre los niveles de detalle dentro de una dimensión. Los hechos son medidas cuantitativas que se almacenan en la base de datos. Los cubos combinan dimensiones y hechos para permitir el análisis desde diferentes perspectivas. Las relaciones se establecen entre dimensiones y hechos para mostrar cómo los datos en una dimensión se relacionan con los hechos.

El diseño lógico proporciona una visión estructural clara de la base de datos multidimensional antes de abordar los detalles de implementación a nivel de sistema de gestión de bases de datos.

Por último el diseño físico, implica la implementación concreta de una base de datos multidimensional en un sistema de gestión de bases de datos (DBMS) específico. Esto incluye la organización y almacenamiento de los datos en el disco, la estructuración de las tablas, la creación de índices y la optimización del rendimiento para consultas multidimensionales.

En este proceso, se traducen los elementos del modelo multidimensional, como dimensiones, jerarquías, hechos y cubos, en tablas reales en la base de datos. Las tablas de hechos contienen las métricas o medidas, mientras que las tablas de dimensiones contienen los atributos de las dimensiones.

Los índices son esenciales para acelerar las consultas en bases de datos multidimensionales, y pueden ser índices simples o multidimensionales que optimizan la búsqueda en múltiples dimensiones. El particionamiento de tablas se utiliza para administrar grandes conjuntos de datos, dividiendo una tabla en subconjuntos más manejables.

Para mejorar el rendimiento, se pueden utilizar tablas de agregados que almacenan valores precalculados y la compresión de datos para reducir el espacio de almacenamiento. Además, se requiere la creación de consultas eficientes que aprovechen la estructura multidimensional de la base de datos.

El diseño físico debe ser monitoreado y ajustado con el tiempo para garantizar un rendimiento óptimo, ya que la base de datos puede crecer y cambiar.

Si desnormalizamos las tablas de dimensiones o no, tendré un esquema de estrella (star) o copo de nieve. Kimball recomienda utilizar siempre la desnormalización total, pero está claro que hay situaciones en las que no queda más opción que pasarnos al esquema copo de nieve (aunque solo sea para alguna dimensión).

El método de diseño elegido es el de Kimball debido a que en un escenario donde el énfasis está en la creación de un Data Warehouse con un enfoque en consultas analíticas y generación de informes, el método de diseño de Kimball es más adecuado que Inmon. Como mencioné previamente, el diseño a utilizar por el método elegido es el de estrella ya que ofrece una estructura simple, desnormalizada y fácil de entender para organizar los datos.

En el diseño de estrella, tengo una tabla central llamada "tabla de hechos" que alberga las métricas y medidas cuantitativas clave que son fundamentales para la toma de decisiones. Además, las tablas de dimensiones contienen atributos descriptivos que permiten el desglose y análisis de los datos en la tabla de hechos, lo que facilita la generación de informes y análisis detallados.

Asimismo también elijo Kimball porque su arquitectura se considera como una aproximación “bottom-up” del problema, ya que partimos de los datos y procesos existentes y modelamos el Data Warehouse para que se adapte a ellos. También el enfoque de Kimball se ajusta más a proyectos pequeños en los que se persigue un sistema fácilmente explotable y entendible por el usuario y de rápido desarrollo.

Este es mi caso ya que parto de la base de datos de suicidios a nivel nacional y a partir de estos datos armo el Data Warehouse adaptado a ellos, al ser un informe de un trabajo práctico integrador, es un proyecto pequeño de rápido desarrollo. Cabe destacar que Kimball no requiere un alto grado de especialización, por lo que se ajusta a los requerimientos del informe ya que no dispongo del tiempo para ser una especialista en el tema.

# Diseño e implementación

Los datos son extraídos de una base de datos pública nacional, junto a la base de datos en un archivo csv también hay un manual de usuario de cómo debe ser utilizada y descripción de las variables. Es por esto que del total de columnas del dataset solo utilizaré las que aporten a la realización de este informe, estas son:

1. id\_hecho
2. provincia\_nombre
3. departamento\_nombre
4. localidad\_nombre
5. fecha\_hecho
6. hora\_hecho
7. tipo\_lugar
8. modalidad
9. suicida\_sexo
10. suicida\_tr\_edad
11. suicida\_identidad\_genero

En cuanto a la columna “suicida\_identidad\_genero” al final no se utilizará ya que es una columna que se agregó en el último año de los registros y según el manual de usuario del dataset “Esta variable se incorporó al SNIC en el año 2021 en el módulo SAT Homicidios Dolosos y SAT Suicidios. Por lo tanto, para los años 2017-2020 no se cuenta con información para esta variable. Disponibilidad del dato: 100%, corresponde a Sin determinar el 78,3% del total de víctimas." Es por esto que esta variable será eliminada y no se tendrá en cuenta.

En el Anexo de este documento está el diccionario de las variables donde se puede ver con más detalle las características de cada una.

Teniendo en cuenta estas columnas del dataset puedo plantear las dimensiones:

1. Edad: Permite analizar la distribución de suicidios por grupos de edad.
2. Sexo: Permite analizar la distribución de suicidios por su sexo.
3. Lugar: Proporciona información sobre la ubicación geográfica de los suicidios.
   * localidad
   * departamento
   * provincia
4. Tiempo: Permite analizar las tendencias temporales y estacionales de los suicidios.
   * fecha
   * hora
   * dia
   * mes
   * anio
5. Modalidad: Describe cómo se llevaron a cabo los suicidios, incluye información sobre los métodos utilizados.
6. Tabla de hechos “Sucesos”:
   * Identificador del hecho
   * claves foráneas de:
     + tiempo
     + lugar
     + sexo
     + edad
     + modalidad

* cantidad de suicidios

Planteo un diagrama de estrella, en mi caso tengo una tabla de hechos denominada “sucesos”. A partir de esto, la métrica que tengo es la “cantidad\_suicidios” que es la única medida ya que es el único hecho que tengo, todas las dimensiones aportan a esta métrica.

Asimismo en la tabla de hechos puedo realizar operaciones OLAP para calcular las métricas derivadas, dichas operaciones se describen en la sección “Operaciones OLAP”.

El diseño multidimensional sin determinar cuales son las claves foráneas o primarias, con el cubo de hechos diferenciado es visible en la imagen número 1.

|  |
| --- |
| *Imagen 1: diagrama multidimensional para base de datos multidimensional de suicidios en Argentina* |

Teniendo en mente que el modelo elegido fue Kimball, las dimensiones y la tabla de hecho con sus métricas, procedo a utilizar Draw.io para graficar los diagramas. Será de tipo estrella ya que es más simple y directo para representar las dimensiones, asimismo las tablas de dimensiones están desnormalizadas para mejorar el rendimiento de las consultas.

Cabe destacar que en la tabla tiempo, la fecha y hora son modificadas mediante un proceso de ETL, que se describirá luego, y se concatenan en la variable fecha. El diagrama de tablas lo podemos ver en la imágen número 2.

|  |
| --- |
| *Imagen 2: diagrama tipo estrella para base de datos multidimensional de suicidios en Argentina* |

Lo que cambia en el modelo multidimensional y en el de tablas, es que en el multidimensional no vemos las claves foráneas en la tabla de hechos, solo vemos la métrica y las métricas derivadas (aquellas que tienen “/” por delante), mientras que en el diagrama de tablas tengo la métrica de la cantidad de suicidios y las claves foráneas de las tablas que aportan a la tabla de hechos.

Con respecto a las funciones derivadas, no se escriben por mes, año o periodo (usando el tiempo como ejemplo) ya que eso lo aclaro en el dashboard mediante operaciones OLAP, ya que la medida derivada indica que voy a hacer con esos datos cuando tenga que agrupar por modalidad, edad, sexo, entre otras, el agregado va a ser una suma “cantidad\_total” que es una suma de la cantidada de suicidios. Se puede calcular con las otras a medida que hago un drill down por ejemplo.

En el caso de la base de datos de suicidios, tengo una tabla de hechos que contiene datos sobre los suicidios y dimensiones como la edad, sexo, lugar, tiempo y modalidad. Cada dimensión se mantendrá en su propia tabla, y estas tablas se conectan directamente con la tabla de hechos a través de claves foráneas dependiendo cada caso. Cada dimensión en este modelo es una tabla independiente y no hay una subdivisión adicional de las dimensiones, lo que facilita la comprensión y el rendimiento de las consultas.

En mi caso parto de un archivo CSV, que contiene los datos de mi base de datos, a partir de esto realizó un proceso de ETL y luego el Data Warehouse, este último utilizaré para hacer los reportes.

Lo primero que debo realizar es el proceso de DDL para crear las tablas en la base de datos que se llama “bd\_suicidios”. A partir de este proceso creo las tablas, asigno sus claves primarias y foráneas, mediante esta acción creo el Data Warehouse vacío.

## ETL

El proceso de Extracción, Transformación y Carga (ETL) es la base sobre la cual se alimenta el Datawarehouse. En mi caso mis datos los extraigo de la base de datos “Hechos y víctimas de suicidios en la República Argentina. Total nacional” que es una base de datos en formato CSV que contiene microdatos relacionados con eventos de suicidio y las personas afectadas en la República Argentina. Estos datos son recopilados a partir del Sistema de Alerta Temprana del Sistema Nacional de Información Criminal y cubren el período a partir de 2017 hasta el 2022.

Para la transformación, lo primero que realizo es mediante una serie de líneas de código, utilizando la librería de Pandas de Python, conectarme a la base de datos en Postgres, leo el archivo CSV y creo un Data Frame “casos\_suicidios” que es el que utilizaré para los pasos siguientes, donde extraigo la información de las columnas del CSV correspondientes a las tablas de la base de datos.

Luego comienzo a cargar las dimensiones, para esto primero creo un Data Frame para cada una y con sus datos correspondientes, luego los cargo utilizando una función provista por la cátedra que “actualiza una tabla de dimensión de un DW con los datos nuevos. Si los datos

ya existen en la tabla, no se agregan. Devuelve la tabla actualizada con los pk tal cual está en la base de datos.” Las tablas de modalidad, edad y sexo al solo tener el ID y una variable descriptiva son las primeras.

Para la dimensión de Lugar, creo el Data Frame, agrego las filas correspondientes a los nombres de las localidades, departamentos y provincias, luego elimino los registros duplicados y por último actualizo la tabla de dimensiones.

Con respecto a la dimensión de Tiempo creo el Data Frame concatenando la fecha del hecho junto con la hora del hecho ya que ambos son de tipo texto, luego transformo estos datos en tipo Datetime para poder trabajar sobre ellos, dándoles el formato de “%d-%m-%Y %H:%M:%S”. Con la carga de los datos en la tabla “Tiempo” tuve inconvenientes que fueron solucionados corrigiendo la función provista por la cátedra de “actualizarTablaDimension”. Para cada operación debo agregar una nueva columna, por ende tengo una columna para:

* Días de la semana: donde 0 es lunes, 1 es martes y así sucesivamente hasta el 6 que corresponde al domingo
* Meses del año: donde 0 es enero, 1 es febrero y así sucesivamente hasta el 11 que corresponde a diciembre
* Año: desde 2017 a 2022
* Periodo: es una variable creada que clasifica las horas en:
  + madrugada (0 a 6am)
  + mañana (6 a 12 am)
  + tarde (12 a 18 pm)
  + noche (18 a 24 pm)

Por último para las tablas de hechos que al igual que las otras tablas, creo un Data Frame denominado “sucesos” donde cargo las dimensiones y la métrica definida previamente con sus respectivos datos. El proceso de carga se realizó en dos etapas, primero se cargan los datos por dimensión y luego se cargan los datos a la tabla de hechos que corresponde.

Antes de cargar los datos a mi tabla de hechos, tuve que separar el data frame “casos\_suicidios” en una variable denominada “agrupado” ya que para realizar el conteo necesitaba agrupar por todas las columnas que le indico y contar cuántos casos hay en cada grupo. Luego comienzo a trabajar sobre “agrupado”, utilizando la función de Python “reset.index(‘counts’)” que me devuelve un dataframe con los valores de la columna "0" y los índices de las columnas que le pase como parámetro, luego utilizó únicamente la columna “counts” que es la columna que tiene la cantidad de suicidios por cada caso. Podemos ver cómo funciona “agrupado” en la imagen número 3.

|  |
| --- |
| *Imagen 3: variable “agrupado”* |

Cabe destacar que nos encontramos frente a casos donde el conteo da como resultado 2, por ende asumimos que para ese día, a esa hora, en ese lugar, dos personas del mismo sexo y en el mismo rango de edad decidieron quitarse la vida.

Para el cálculo de la métrica “cantidad\_suicidios” se trabajó sobre la variable agrupado, ya que teníamos la variable conteo, la métrica solo llama a esta.

Por la simplicidad del código, considero que es lo suficientemente robusto para que funcione con todos los datos y flexible para que cuando haya una modificación en el ETL o en la fuente de datos, este no se arruine ya que contempla las dimensiones más importantes a la hora de calcular la métrica y sus derivas, además de poder extraer información valiosa sobre los suicidios a nivel nacional.

## Operaciones OLAP

Con un modelo multidimensional que organiza los datos en dimensiones y medidas, el OLAP permite consultas avanzadas, análisis ad hoc y la generación de informes, lo que resulta fundamental para la toma de decisiones estratégicas. Además, el OLAP optimiza el rendimiento en consultas analíticas, almacena datos históricos y proporciona una visión más intuitiva de los datos, lo que resulta crucial en diversas industrias y sectores.

Asimismo hay tres aproximaciones dependiendo de cómo son almacenados los datos en el cubo:

1. MOLAP: El procesamiento analítico multidimensional en línea (MOLAP) implica la creación de un cubo que representa datos multidimensionales de un almacenamiento de datos. El sistema MOLAP almacena en el hipercubo los datos calculados previamente. Se utiliza MOLAP porque este tipo de tecnología OLAP proporciona un análisis rápido.
2. ROLAP: En lugar de utilizar un cubo de datos, el procesamiento analítico relacional en línea (ROLAP) permite hacer análisis de datos multidimensionales en una base de datos relacional. En otras palabras, utilizan consultas SQL para buscar y recuperar información específica según las dimensiones requeridas. ROLAP es adecuado para analizar datos extensos y detallados. Sin embargo, ROLAP tiene un rendimiento de consultas lento en comparación con MOLAP.
3. HOLAP: El procesamiento analítico híbrido en línea (HOLAP) combina MOLAP y ROLAP para ofrecer lo mejor de ambas arquitecturas. HOLAP permite recuperar rápidamente los resultados analíticos de un cubo de datos y extraer información detallada de las bases de datos relacionales.

Por la naturaleza de la base de datos de suicidios opté por utilizar ROLAP ya que es útil cuando hay un gran volumen de datos o datos almacenados en bases de datos relacionales existentes. Es flexible y puede manejar conjuntos de datos más grandes. Adecuado cuando las necesidades de consulta son complejas y se requiere acceso a datos en tiempo real.

Para la tabla “Sucesos”:

1. Corte (Slicing): permite observar datos para un valor específico en una dimensión, es decir, puedo cortar los datos por sexo para analizar la distribución de suicidios por sexo.
2. Drilling Down/Up: permite obtener más detalles, podría comenzar viendo la distribución de suicidios por año y, luego, perforar hacia abajo para ver la distribución mensual o diaria.
3. Distribución (Dicing): permite enfocar en una combinación de valores en varias dimensiones, es decir, puedo analizar la distribución de suicidios por sexo y lugar, para comprender cómo varía por ubicación geográfica.
4. Agregación (Roll Up): permite observar datos a un nivel más alto de resumen. Podría agregar los datos para ver la distribución de suicidios en una provincia en lugar de una localidad específica.
5. Pivoting: permite comparar los datos entre dos o más dimensiones. Por ejemplo, podría comparar la distribución de suicidios por género y edad en diferentes provincias.

## Plan de mantenimiento

Aunque a nivel nacional la base de datos con la que se trabaja tiene una actualización anual, se propone que el mantenimiento se realice de forma cuatrimestral, para que la información esté actualizada y se puedan prevenir el error frente a la carga de datos cuando el volumen es muy grande, dado que estamos hablando de un listado de suicidios a nivel nacional.

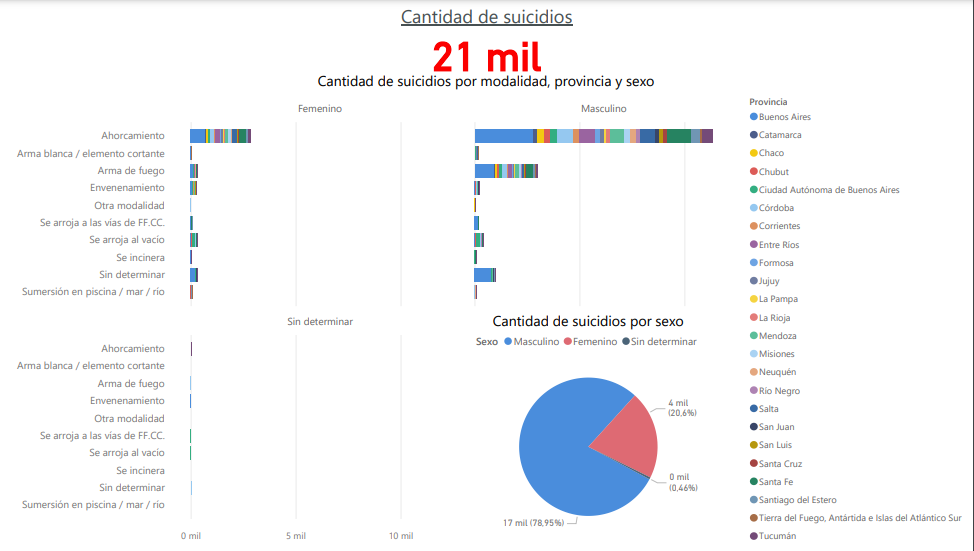
Los pasos para el mantenimiento se resumen en tres variables “cómo”, “cuándo” y el “contexto” frente al cual se da la acción.

* Cómo: se actualizará mediante un proceso de ETL.
* Cuándo: se propone que sea de forma cuatrimestral, para tener información real de lo que está sucediendo a nivel nacional y cómo ayudar a las personas que deciden finalizar con su vida. Para poner más énfasis según fechas y lugares que lo requieran e intentar prevenir. De forma de Batch periódico ya se que el mantenimiento se realiza en intervalos regulares de tiempo, es decir, cada cuatro meses.
* Contexto: se realizará fuera de línea, por ende no pueden realizarse transacciones de mantenimiento y consulta simultáneamente, es decir, en este contexto el Data Warehouse está inhabilitado a los analistas de la organización ya que se está actualizando y no pueden realizarse consultas y/o actualizaciones al mismo tiempo.

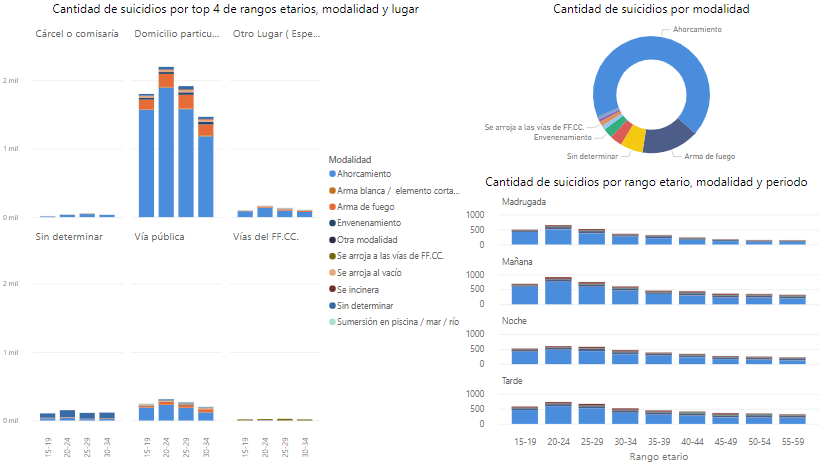
## Reportes

Se realizó una conexión entre PowerBi y la base de datos multidimensional para poder realizar un dashboard para visualizar la información de la métrica principal “cantidad\_suicidios” y sus derivadas.

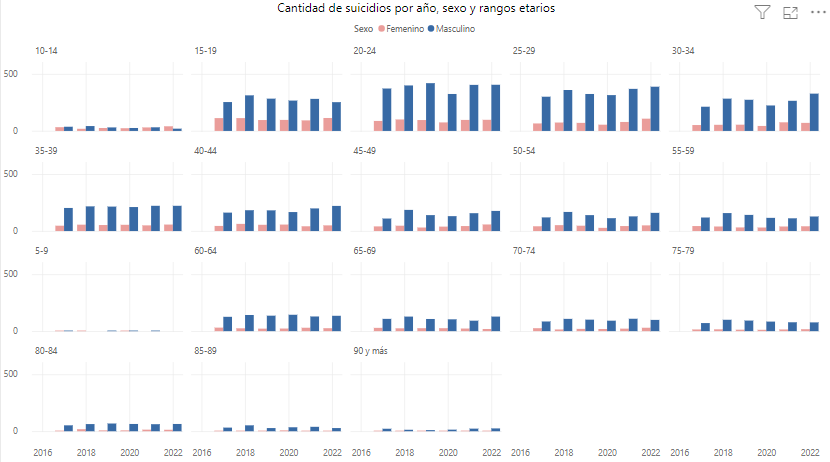
Se utilizaron distintas dimensiones para cada gráfico, para poder visualizar la información de manera más simple y directa posible. Se optó por utilizar los nombres de las modalidades, rangos etarios, provincias y tipos de lugares a fines de que sea más claro el gráfico y no estar recurriendo a los identificadores.



En el primer dashboard podemos ver cómo se comportan las distintas modalidades de suicidios según cada provincia y el sexo de la persona. Asimismo vemos que la distribución de hombres es mucho mayor a las de mujeres, 78% y 20% respectivamente. Las provincias que más suicidios tienen son Buenos Aires, Santa Fe, Córdoba y Entre Ríos.



En el segundo dashboard primero vemos según la edad cual es el lugar donde más suicidios ocurren independientemente de la provincia, que es el domicilio particular. La modalidad más recurrente es el ahorcamiento y luego por arma de fuego. En cuanto al período, las personas tienden a finalizar su vida de mañana y de tarde.



En el tercer dashboard vemos como se distribuyen los suicidios a lo largo de los años, según el rango etario de la persona y su sexo. Otra vez volvemos a notar la predominancia del sexo amsculino ante el femenino, dentro de todos los rangos. Vemos que en el año 2020 los suicidios disminuyen frente al 2019 que tuvieron un pico. El rango etario en el que más personas deciden terminar con su vida es el de 20 a 24 años, siguiéndole 25 a 29 y por último 15 a 19 años.

Cabe aclarar que los rangos de edad se toman desde los 5 a 9 y finalizan en 90 y más. En todos los rangos hay suicidios contabilizados.

# Conclusión

Es posible prevenir los suicidios adoptando medidas a nivel de la población, de determinados grupos poblacionales y del individuo.

Para finalizar, podemos afirmar que respondimos la interrogante inicial que era “¿Cuál sería la metodología adecuada para examinar los patrones y tendencias de la base de datos de suicidios a nivel nacional, con el propósito de informar el diseño de una nueva campaña de prevención del suicidio que será implementada por el Ministerio de Salud y el Ministerio de Seguridad en el período comprendido entre 2017 y 2022?” Ya que se diseñó un datawarehouse que incluye las dimensiones principales para analizar las tendencias y patrones a lo largo de nuestro país.

A partir de los resultados obtenidos notamos que independientemente de la modalidad, lugar, y rango etario, las personas de sexo masculino tienden en mayor medida a terminar con su vida ante las personas de sexo femenino. Si nos enfocamos en el rango etario, las personas de 20 a 24 años están en el primer puesto y la modalidad más elegida independientemente de la edad es el ahorcamiento en los períodos de la mañana o la tarde. En cuanto a lugares la gran mayoría opta por terminar su vida en su domicilio particular, en cuanto a provincias las que más suicidios tienen son Buenos Aires, Santa Fe, Córdoba y Entre Ríos.

Las actividades preventivas exigen la coordinación y colaboración de varios sectores de la sociedad, incluidos la salud, educación, justicia, política, medios de comunicación, entre otros. Esas actividades deben ser amplias e integrales, dado que ningún enfoque puede atajar por sí solo una cuestión tan compleja.

A nivel personal, por falta de tiempo queda a trabajo futuro evaluar aquellas variables que a fines de este informe no se tuvieron en cuenta como los métodos que originaron la llamada o el género de la persona. Asimismo aplicar el sistema de indexación para el sistema de mantenimiento del datawarehouse y el anexo con aplicaciones para que el público en general pueda interactuar con los gráficos y estadísticas, y los datos en crudo, para poder llevar este tema tan importante a todos.

# Anexo

## Diccionario de datos

En esta sección se presenta la descripción detallada de las variables (o campos) que contiene la base usuaria SAT-SS.

**1. id\_hecho:**

Nombre de la variable: id\_hecho

Descripción: Código de identificación del hecho.

Tipo de variable: Numérica

Comentarios y advertencias: Permite identificar a las personas involucradas en un mismo hecho: víctimas e inculpados. Al ser identificador no contiene valores perdidos. A partir de esta variable se pueden calcular la cantidad de hechos y de víctimas.

**2. provincia\_nombre**

Nombre de la variable: provincia\_nombre

Descripción: Provincia donde ocurrió el hecho

Tipo de variable:Texto

Comentarios y advertencia: Disponibilidad del dato: 100% (no contiene valores perdidos).

**3. departamento\_nombre**

Nombre de la variable: departamento\_nombre

Descripción: Nombre del departamento geográfico donde ocurrió el hecho.

Tipo de variable: Texto

**4. localidad\_nombre**

Nombre de la variable: localidad\_nombre

Descripción: Nombre de la localidad donde ocurrió el hecho

Tipo de variable: Texto

Comentarios y advertencias: En general, las Fuerzas Federales registran sus datos en la categoría Localidad sin determinar, ya que la información se reporta desde sus unidades operativas cuyo alcance territorial no coincide con la división política de cada jurisdicción. Disponibilidad del dato: 100%. Del total de hechos, 0,2% se encuentra en la categoría Localidad sin determinar.

Validación del dato: en la carga de la localidad está condicionada por el departamento y provincia cargados previamente.

**5. fecha\_hecho**

Nombre de la variable: fecha\_hecho

Descripción: Fecha de ocurrencia del hecho

Tipo de variable: Fecha. Formato fecha: dd/mm/aaaa.

Comentarios y advertencias: Se valida con las variables anio y mes. Disponibilidad del dato 100%.

**6. hora\_hecho**

Nombre de la variable: hora\_hecho

Descripción: Hora del hecho

Tipo de variable: Texto

Valores con los que aparece: Formato hora: hh:mm:ss , desde 00:00:00 a 23:59:59.

**7. tipo\_lugar**

Nombre de la variable: tipo\_lugar

Descripción: Lugar de ocurrencia del hecho

Tipo de variable: Categórica

Valores con los que aparece:

| Código | Descripción |
| --- | --- |
| 1 | Vía pública |
| 2 | Domicilio particular |
| 3 | Vías del ferrocarril |
| 4 | Cárcel o comisaría |
| 5 | Otro lugar |
| 99 | Sin Determinar |

**8. modalidad**

Nombre de la variable: modalidad

Descripción: Modalidad utilizada para llevar a cabo el acto

Tipo de variable: Categórica

Valores con los que aparece:

| Código | Descripción |
| --- | --- |
| 1 | Arma de fuego |
| 2 | Arma blanca/elemento cortante |
| 3 | Sumersión piscina/ mar/ río |
| 4 | Envenenamiento |
| 5 | Ahorcamiento |
| 6 | Se arroja al vacío |
| 7 | Se arroja a las vías del ferrocarril |
| 8 | Otra modalidad |
| 9 | Se incinera |
| 99 | Sin determinar |

**9. suicida\_sexo**

Nombre de la variable: sexo\_suicida

Descripción: Sexo del suicida

Tipo de variable: Categórica.

Valores con los que aparece:

| Código | Descripción |
| --- | --- |
| 1 | Femenino |
| 2 | Masculino |
| 99 | Sin determinar |

Comentarios y advertencias: Disponibilidad del dato 100%. Corresponden a Sin determinar el 0.4% del total de víctimas.

**10. suicida\_tr\_edad**

Nombre de la variable: tr\_edad\_suicida

Descripción: Edad del suicida en tramos

Tipo de variable: Categórica

Valores con los que aparece:

| Código | Descripción |
| --- | --- |
| -1 | Sin determinar |
| 1 | 4 o menos |
| 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 | 85-89 |
| 19 | más de 90 |