Imagen que contiene Forma

Descripción generada automáticamente

**PROYECTO AGORA**

**ENTENDIMIENTO DE LOS DATOS - SIVIGILA**

**DOCUMENTO INICIAL**

**2024**

**Control de cambios**

|  |  |  |  |
| --- | --- | --- | --- |
| **VERSIÓN** | **FECHA** | **MODIFICACIÓN** | **RESPONSABLE** |
| 1.3 | 26 de abril de 2024 | Documento inicial | Luis Carlos Chicaíza |
| 1.5 | 08 de mayo de 2024 | Se realiza modificaciones al documento con base a los comentarios recibidos. | Luis Carlos Chicaíza |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Contenido

[1. Introducción 5](#_Toc165389886)

[2. Metodología 6](#_Toc165389887)

[2.1 Exploración de los datos 6](#_Toc165389888)

[3. Fuente principal de los datos 7](#_Toc165389889)

[3.1 Origen y procedencia de los datos 7](#_Toc165389890)

[3.2 Descripción general de SIVIGILA 8](#_Toc165389891)

[3.2.1 Ajustes requeridos en nombres de las fuentes 9](#_Toc165389892)

[3.2.2 Ajustes requeridos en registros de resumen en la exportación 9](#_Toc165389893)

[3.3 Estadísticas descriptivas de las variables 11](#_Toc165389894)

[3.3.1 Esquema para la exploración y análisis descriptivo de los datos 12](#_Toc165389895)

[3.3.2. Listado detallado de la fuente 13](#_Toc165389896)

[3.3.3 Análisis preliminares 66](#_Toc165389897)

[3.4 Análisis de correlación 67](#_Toc165389898)

[3.5 Visualización de datos 68](#_Toc165389899)

[2.5.1 Variable: Edad 68](#_Toc165389900)

[3.5.2 Variable: SemanaEpidemilogica 69](#_Toc165389901)

Ilustraciones

[Imagen 1. Carpetas por años de la fuente SIVIGILA 7](#_Toc165378041)

[Imagen 2. Valores no numéricos variable Edad 9](#_Toc165378042)

[Imagen 3. Nuevos datos variable Edad 10](#_Toc165378043)

[Imagen 4. Visualización de los datos de la variable Edad 68](#_Toc165378044)

[Imagen 5. Visualización de los datos de la variable SemanaEpidemilogica 69](#_Toc165378045)

## Introducción

El presente documento ofrece una visión detallada de los principales hallazgos derivados del proceso de exploración inicial de los datos del sistema SIVIGILA. En esta fase inicial de análisis, se identificó la fuente de datos y se examinaron las variables contenidas en ella y sus características fundamentales.

La fuente de datos denominada SIVIGILA, comprende información recopilada durante el periodo comprendido entre 2008 a 2022. Este periodo permite analizar una variedad de contextos y condiciones que resultan ser importantes para la comprensión y el análisis de las tendencias relacionadas.

Durante la fase inicial de exploración, se realizó un análisis exhaustivo de las variables contenidas en el conjunto de datos del sistema SIVIGILA. Se aplicaron técnicas estadísticas descriptivas básicas para examinar la distribución, la variabilidad y las tendencias presentes en los registros disponibles. Este análisis preliminar sirvió como punto de partida para comprender mejor la naturaleza de los datos y así poder identificar posibles áreas de interés y enfoque para análisis posteriores.

En estas secciones, se presenta la metodología de exploración de los datos, como los principales resultados obtenidos en esta fase inicial, incluyendo las características de las variables analizadas, las tendencias y patrones identificados en los datos del sistema SIVIGILA.

## Metodología

Con el propósito de realizar un análisis exhaustivo de los datos provenientes de la fuente SIVIGILA proporcionada, se busca aplicar diversas técnicas de análisis exploratorio de los datos. Algunas técnicas son las siguientes:

### 2.1 Exploración de los datos

**Análisis de estadísticas descriptivas**

Calcular la media, mediana, desviación estándar y los percentiles, para así comprender la distribución y la variabilidad de las variables numéricas del conjunto de datos.

**Análisis de correlación**

Explorar la relación entre pares de variables numéricas mediante el cálculo de coeficientes de correlación, como el coeficiente de correlación de Pearson. Con el objetivo de identificar posibles asociaciones lineales o no lineales entre las variables.

**Análisis de frecuencias**

Examinar la distribución de valores únicos en variables categóricas para comprender la proporción de cada categoría y detectar posibles desequilibrios o sesgos en los datos.

**Visualización de datos**

Generar gráficos y visualizaciones, como histogramas, diagramas de dispersión y gráficos de barras, para representar gráficamente la distribución de los datos y las relaciones entre variables.

# Fuente principal de los datos

## 3.1 Origen y procedencia de los datos

A continuación, se lista los datos asociados a la base de datos recibida para el proyecto AGORA:

|  |  |
| --- | --- |
| Nombre de la fuente en el proyecto: | SIVIGILA |
| Número de archivos o tablas1: | 15 carpetas con la información por año |
| Nombre del archivo original de la fuente entregado2: | sivigila.zip |
| Ruta del archivo original y archivos de trabajo3: | //caoba-access.virtual. .dfs.core.windows.net/sivigila/sivigila\_parquet\_{año} |

Tabla 1. Descripción inicial de la fuente SIVIGILA

Para el procesamiento de los datos, se definió unos contendores en la infraestructura con la información por año. Esta información está consolidada desde el año 2008 hasta el año 2022.

Imagen 1. Carpetas por años de la fuente SIVIGILA

El procesamiento se realizará sobre estas carpetas para poder analizar de una fecha específica o consolidar todas las fechas para un mayor análisis.

## 3.2 Descripción general de SIVIGILA

A continuación, se detalla los registros por año de los registros en la fuente:

|  |  |  |  |
| --- | --- | --- | --- |
| Base | #  de Registros | # de Columnas | ID Únicos |
| SIVIGILA \_2008\_2011 | 1.513.282 | 57 | 1.153.913 |
| SIVIGILA\_2012\_2020 | 5.877.748 | 57 | 4.743.009 |
| SIVIGILA\_2021\_2023 | 1.445.669 | 57 | 1.181.537 |

Tabla 2. Descripción inicial de la fuente SIVIGILA

El total de registros es de 7’815.255 registros en la fuente SIVIGILA.

### 3.2.1 Ajustes requeridos en nombres de las fuentes

La información recibida abarca un periodo que comprende desde 2008 hasta 2022. No fue necesario aplicar ajuste en los nombres de las variables, no se evidencio un numero o nombre distinto de las columnas a través del tiempo. También la información recibida no evidencia registros iguales o duplicados. Sin embargo, si se presenta una alteración de los datos recibidos, puesto que se encuentran datos atípicos, probablemente correspondientes a otras consultas.

### 3.2.2 Ajustes requeridos en registros de resumen en la exportación

Al analizar los valores no numéricos de los datos originales de SIVIGILA se identificó que la columna Edad estaba compuesta por datos categóricos:

Texto

Descripción generada automáticamente

Imagen 2. Valores no numéricos variable Edad

Por lo cual se hizo la modificación de esta columna para extraer los valores numéricos correspondientes a la edad y así poder generar el análisis estadístico correspondiente.

Tabla

Descripción generada automáticamente

Imagen 3. Nuevos datos variable Edad

Una vez modificada, se procedió a eliminar los campos con valores en NULL.

## 3.3 Estadísticas descriptivas de las variables

En esta fase se realizará el análisis exploratorio de los datos de la fuente SIVIGILA. Esta fase busca obtener una compresión más profunda de la naturaleza y las características básicas de las variables que componen este conjunto de datos.

A través de algunas técnicas de estadística y visualizaciones, se explorará la distribución, la variabilidad y las posibles relaciones entre las diferentes variables. También, se buscará identificar algunos patrones, tendencias y anomalías que puedan proporcionar información importante para futuros análisis y la toma de decisiones.

Este análisis permitirá generar una contextualización adecuada de los datos, y sentar las bases para la limpieza de los datos adecuada para el entorno de la investigación y realizar análisis mucho más robustos en etapas posteriores a este proyecto

### 3.3.1 Esquema para la exploración y análisis descriptivo de los datos

A continuación, se define el siguiente esquema para realizar la exploración y el análisis descriptivo de los datos. Para cada variable de la fuente de datos SIVIGILA se debe diligenciar la siguiente tabla.

|  |  |
| --- | --- |
| **FUENTE** | **Nombre de la fuente / Nombre del archivo o tabla** |
| **Campo 1** | Nombre del campo tal como aparece en el archivo fuente |
| Descripción | Tipo de dato (fecha, texto categórico, texto no categórico, numérico nominal o numérico ordinal.) |
| Periodo | Periodo de la base de datos |
| Extensión | Formato del archivo |
| Volumen (#filas) | Cantidad de registros |
| Campos (#columnas) | Numero de atributos |
| **Campos SIVIGILA** | Descripción de todos los atributos de la base de datos |
| **Campo n** | nombre del campo en la fuente de datos |
| Alias | Nombre de la variable |
| Descripción | Descripción o propósito del atributo. |
| Tipo | Tipo de dato (fecha, texto categórico, texto no categórico, numérico nominal o numérico ordinal.) |
| Dominio | Rango de los datos o valor esperado. |
| Estadísticas básicas | Se calculan las estadísticas básicas o la frecuencia de las variables. |
| Observaciones | Información relevante de la variable a tener en cuenta |

*Tabla 3: Descripción estadística de las variables*​

### 3.3.2. Listado detallado de la fuente

Teniendo en cuenta el esquema mencionado anteriormente, se genera un cuadro detallado describiendo la fuente de datos y la descripción de cada una de las variables identificadas.

|  |  |
| --- | --- |
| **FUENTE** | **INFORMACIÓN SIVIGILA** |
| Descripción | Contiene los datos del sistema SIVIGILA |
| Período | 2008 – 2022 |
| Extensión | .parquet |
| Volumen (#filas) | 8’836.699 7’815.255 |
| Campos (#columnas) | 57 |
| Fecha recepción | 7 de diciembre de 2023 |
| **Campo 1** | PersonaBasicaID |
| Alias | ID de la persona |
| Descripción | Corresponde al identificador de la persona anonimizado. |
| Tipo | Categórico Nominal |
| Dominio | [ENTERO POSITIVO]  Existen 6’064.983 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * El valor más frecuente es 1, con una cuenta de 685,388 registros. |
| **Campo 2** | Evento |
| Alias | Describe el evento. |
| Descripción | Define el tipo del evento con su código  asociado. |
| Tipo | Categórico Nominal |
| Dominio | [Código, Nombre de evento]  Existen 79 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Los eventos más comunes son agresiones (código 300), varicela (código 831), y dengue (código 210), esto podría indicar que estos eventos son prevalentes o que hay una alta incidencia de estos eventos en la población analizada. |
| **Campo 3** | AñoEpidemiologico |
| Alias | Año epidemiológico. |
| Descripción | Define el año epidemiológico. |
| Tipo | Fecha |
| Dominio | [2008-2022]  Existen 15 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Las fechas corresponden a el rango de fecha proporcionado. |
| **Campo 4** | SemanaEpidemilogica |
| Alias | Semana epidemiológica |
| Descripción | Define la semana epidemiológica. |
| Tipo | Numérico |
| Dominio | [01 a 53] |
| Estadísticas Básicas |  |
| Observaciones | * La media correspondiente a semanas epidmiologicas corresponde a 26.4 |
| **Campo 5** | Prestador |
| Alias | Prestador |
| Descripción | Nombre de los prestadores de servicios. |
| Tipo | Categórico Nominal. |
| Dominio | [Código - Nombre del prestador]  Existen 11.488 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * El prestador más común es "0 - NO DEFINIDO", otros prestadores tienen cuentas significativamente menores, |
| **Campo 6** | CodigoAdministradora |
| Alias | Código de la administradora. |
| Descripción | Proporciona el código asociado a la administradora. |
| Tipo | Categórico Nominal. |
| Dominio | [Código alfanumérico]  Existen 419 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los casos están relacionados con "NO DEFINIDO" o con códigos específicos como EPS002, EPS037, EPS01 en mayor proporción. |
| **Campo 7** | Administradora |
| Alias | Administradora |
| Descripción | Define el nombre de la administradora |
| Tipo | Categórico Nominal |
| Dominio | [SALUD, NO DEFINIDO, ASEGURADORAS, GUAVIARE]  Existen 4 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los casos (6,613,148) están relacionados con "SALUD". * Las demás categorías (NO DEFINIDO, ASEGURADORAS, GUAVIARE) tienen un número considerablemente menor de casos. |
| **Campo 8** | Edad |
| Alias | Edad |
| Descripción | Define la edad del usuario o paciente. |
| Tipo | Numérico |
| Dominio | [ 1 a 95000] |
| Estadísticas Básicas |  |
| Observaciones | * La media correspondiente a la edad con un valor de 24.5, que puede indicar una mayor nivel de ateción a jovenes. |
| **Campo 9** | UnidadMedidaEdad |
| Alias | Unidad Medida Edad |
| Descripción | Describe la unidad de medida para la edad |
| Tipo | Categórico Nominal. |
| Dominio | [AÑOS, MESES, DÍAS, SALUD, HORAS, MINUTOS]  Existen 6 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Los valores de "AÑOS" son los más frecuentes, representando la mayoría de los datos. |
| **Campo 10** | Sexo |
| Alias | Sexo |
| Descripción | Define el Sexo del usuario o paciente. |
| Tipo | Numérico. |
| Dominio | [FEMENINO, MASCULINO, INDEFINIDO]  Existen 105 valores únicos, evidentes 3. |
| Estadísticas Básicas |  |
| Observaciones | * Se puede observar unos datos atipicos correspondientes a la columna de Edad, probablemente debido a un error de carga en los datos. |
| **Campo 11** | TipoRegimen |
| Alias | Tipo Régimen |
| Descripción | Define el tipo de régimen del usuario o paciente. |
| Tipo |  |
| Dominio | [S - SUBSIDIADO, C – CONTRIBUTIVO, N - NO AFILIADO, P – ESPECIAL, E – EXCEPCION, ND - NO DEFINIDO, AÑOS, MESES, DÍAS, NO APLICA]  Existen 10 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los casos se encuentran en los tipos de régimen "SUBSIDIADO" y "CONTRIBUTIVO". * Se evidencia campos de otras columnas. |
| **Campo 12** | codigoOcupacion |
| Alias | Código ocupación |
| Descripción | Define el código de ocupación. |
| Tipo | Numérico |
| Dominio | [ENTERO POSITIVO]  Existen 403 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Se puede analizar que la mayoria de registros estan sin asignar, con el código “ND” |
| **Campo 13** | Ocupacion |
| Alias | Ocupación |
| Descripción | Describe el tipo de ocupación, dado un código y el nombre de esta. |
| Tipo | Categórico Nominal. |
| Dominio | [Código – Nombre ocupación]  Existen 405 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los casos están relacionados con ocupaciones no definidas ("NO DEFINIDO"). * Las siguientes ocupaciones más comunes incluyen "6111 - AGRICULTORES Y TRABAJADORES" y "7111 - MINEROS Y TRABAJADORES". |
| **Campo 14** | Etnia |
| Alias | Etnia |
| Descripción | Describe la etnia del usuario o paciente. |
| Tipo | Categórico Nominal |
| Dominio | Existen 224 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Se pueden observar varios valores atipicos, al parecer correspondiente a otras columnas. |
| **Campo 15** | ClasificacioninicialCaso |
| Alias | Clasificación inicial caso |
| Descripción | Describe la clasificación inicial asociada. |
| Tipo | Categórico Nominal. |
| Dominio | Existen 131 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Se pueden observar varios valores atipicos, al parecer correspondiente a otras columnas. |
| **Campo 16** | PacienteHospitalizado |
| Alias | Paciente hospitalizado |
| Descripción | Describe si el paciente fue hospitalizado. |
| Tipo | Categórico Nominal |
| Dominio | Existen 45 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Se pueden observar varios valores atipicos, al parecer correspondiente a otras columnas. |
| **Campo 17** | SeguimientoFinalCaso |
| Alias | Seguimiento final caso |
| Descripción | Describe el tipo de seguimiento final para el caso. |
| Tipo | Categórico Nominal |
| Dominio | Existen 23 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Se pueden observar varios valores atipicos, al parecer correspondiente a otras columnas. |
| **Campo 18** | FechaNotificacion |
| Alias | Fecha notificación |
| Descripción | Describe la fecha de notificación |
| Tipo | Categórico Nominal / Fecha |
| Dominio | Existen 5.586 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Se pueden observar varios valores atipicos, al parecer correspondiente a otras columnas. |
| **Campo 19** | FechaConsulta |
| Alias | Fecha consulta |
| Descripción | Describe la fecha de consulta |
| Tipo | Categórico Nominal / Fecha |
| Dominio | Existen 5.590 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Se pueden observar varios valores atipicos, al parecer correspondiente a otras columnas. |
| **Campo 20** | FechaInicioSINTOmas |
| Alias | Fecha inicio síntomas |
| Descripción | Describe la fecha de inicio de los síntomas |
| Tipo | Categórico Nominal / Fecha |
| Dominio | Existen 5.895 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Se pueden observar varios valores atipicos, al parecer correspondiente a otras columnas. |
| **Campo 21** | FechaHospitalizacion |
| Alias | Fecha hospitalización |
| Descripción | Describe la fecha de hospitalización |
| Tipo | Categórico Nominal / Fecha |
| Dominio | Existen 8.865 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Se pueden observar varios valores atipicos, al parecer correspondiente a otras columnas. |
| **Campo 22** | FechaDefuncion |
| Alias | Fecha de defunción. |
| Descripción | Describe la fecha de defunción. |
| Tipo | Categórico Nominal / Fecha |
| Dominio | [AAAAMMDD]  Existen 6682 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Se evidencia en mayor numero el registro “19000101” con 2122930 ocurrencias, este un valor predeterminado que puede indicar una fecha desconocida o no especificada. * Se pueden observar varios valores atipicos, al parecer correspondiente a otras columnas. |
| **Campo 23** | FechaNcto |
| Alias | Fecha de Nacimiento. |
| Descripción | Describe la fecha de Nacimiento. |
| Tipo | Categórico Nominal / Fecha |
| Dominio | [AAAAMMDD]  Existen 37.088 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Se pueden observar varios valores atipicos, al parecer correspondiente a otras columnas. * Se evidencia en mayor numero el registro “19000101” con 4074619 ocurrencias, este un valor predeterminado que puede indicar una fecha desconocida o no especificada. |
| **Campo 24** | FechaAjuste |
| Alias | Fecha de ajuste |
| Descripción | Describe la fecha de ajuste. |
| Tipo | Categórico Nominal / Fecha |
| Dominio | [AAAAMMDD]  Existen 21842 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros se encuentran con el valor por defecto. |
| **Campo 25** | GP\_Discapacidad |
| Alias | GP Discapacidad |
| Descripción | Describe la fecha del grupo poblacional en personas con situación de discapacidad |
| Tipo | Categórico Nominal |
| Dominio | [SI, NO]  Existen 38.194 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los casos se encuentra en la categoría "NO". * Existen datos atípicos de fechas correspondientes a otras columnas. |
| **Campo 26** | GP\_Despazados |
| Alias | GP Despazados |
| Descripción | Describe si aplica al grupo poblacional de personas con situación de desplazados. |
| Tipo | Categórico Nominal |
| Dominio | [SI, NO]  Existen 26.122 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los casos se encuentra en la categoría "NO". * Existen datos atípicos de fechas correspondientes a otras columnas. |
| **Campo 27** | GP\_Migrantes |
| Alias | GP migrantes |
| Descripción | Describe si aplica al grupo poblacional de personas con situación de migrantes. |
| Tipo | Categórico Nominal |
| Dominio | [SI, NO]  Existen 32.783 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los casos se encuentra en la categoría "NO". * Existen datos atípicos de fechas correspondientes a otras columnas. |
| **Campo 28** | GP\_Carcelarios |
| Alias | GP Carcelarios |
| Descripción | Describe si aplica al grupo poblacional de personas privadas de la libertad. |
| Tipo | Categórico Nominal |
| Dominio | [SI, NO]  Existen 11.515 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los casos se encuentra en la categoría "NO". * Existen datos atípicos de fechas correspondientes a otras columnas. |
| **Campo 29** | GP\_Gestantes |
| Alias | GP Gestantes |
| Descripción | Describe si aplica al grupo poblacional de gestantes. |
| Tipo | Categórico Nominal |
| Dominio | [SI, NO]  Existen 4956 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los casos se encuentra en la categoría "NO". * Existen datos atípicos de fechas correspondientes a otras columnas. |
| **Campo 30** | GP\_Indigen |
| Alias | GP Indigen |
| Descripción | Describe si aplica al grupo poblacional en personas con situación de Habitantes de Calle |
| Tipo | Categórico Nominal |
| Dominio | [SI, NO]  Existen 1.339 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los casos se encuentra en la categoría "NO". * Existen datos atípicos de fechas correspondientes a otras columnas. |
| **Campo 31** | GP\_PoblacionICBF |
| Alias | GP Población ICBF |
| Descripción | Describe si aplica al grupo poblacional de población infantil a cargo del ICBF |
| Tipo | Categórico Nominal |
| Dominio | [SI, NO] |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los casos se encuentra en la categoría "NO". * Existen datos atípicos de fechas correspondientes a otras columnas. |
| **Campo 32** | GP\_MadresComunitarias |
| Alias | GP Madres comunitarias |
| Descripción | Describe si aplica en el grupo poblacional de madres comunitarias. |
| Tipo | Categórico Nominal |
| Dominio | [SI, NO]  Existen 3 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los casos se encuentra en la categoría "NO". * Existen datos atÍpicos de fecha correspondientes a otra columna. |
| **Campo 33** | GP\_Desmovilizados |
| Alias | GP Desmovilizados |
| Descripción | Describe si aplica al grupo poblacional de desmovilizados. |
| Tipo | Categórico Nominal |
| Dominio | [SI, NO]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los casos se encuentra en la categoría "NO". |
| **Campo 34** | GP\_Psiquia |
| Alias | GP Psiquia |
| Descripción | Describe si aplica al grupo poblacional de población en centros psiquiátricos. |
| Tipo | Categórico Nominal |
| Dominio | [SI, NO]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los casos se encuentra en la categoría "NO". |
| **Campo 35** | GP\_VictimasViolencia |
| Alias | GP Victimas violencia |
| Descripción | Describe si aplica al grupo poblacional de víctimas de la violencia |
| Tipo | Categórico Nominal |
| Dominio | [SI, NO]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los casos se encuentra en la categoría "NO". |
| **Campo 36** | GP\_otros |
| Alias | GP Otros |
| Descripción | Describe si aplica al grupo poblacional catalogada como otros. |
| Tipo | Categórico Nominal |
| Dominio | [SI, NO]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los casos se encuentra en la categoría "NO". |
| **Campo 37** | codigoPaisOcurrencia |
| Alias | Código País Ocurrencia |
| Descripción | Describe el código del país de ocurrencia |
| Tipo | Categórico Nominal |
| Dominio | [SI, NO]  Existen 93 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los casos se encuentra en la categoría "NO", es un valor que corresponde a otra columna. * Existen datos atípicos correspondientes a otras columnas. |
| **Campo 38** | PaisOcurrencia |
| Alias | País ocurrencia |
| Descripción | Define el nombre del país de ocurrencia. |
| Tipo | Categórico Nominal |
| Dominio | Existen 120 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Se pueden observar varios valores atipicos, al parecer correspondiente a otras columnas. |
| **Campo 39** | codigoDepartamentoOcurrencia |
| Alias | Código departamento ocurrencia. |
| Descripción | Describe el código del departamento del evento |
| Tipo | Categórico Nominal |
| Dominio | [Entero positivo]  Existen 176 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Se pueden observar varios valores atipicos, al parecer correspondiente a otras columnas. |
| **Campo 40** | DepartamentoOcurrencia |
| Alias | Departamento Ocurrencia |
| Descripción | Describe el nombre del departamento del evento. |
| Tipo | Categórico Nominal |
| Dominio | Existen 215 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Se pueden observar varios valores atipicos, al parecer correspondiente a otras columnas. |
| **Campo 41** | codigoMunicipioOcurrencia |
| Alias | Código municipio ocurrencia |
| Descripción | Describe el código del municipio del evento según DANE |
| Tipo | Categórico Nominal |
| Dominio | [Entero Positivo]  Existen 1.298 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Se pueden observar varios valores atipicos, al parecer correspondiente a otras columnas. |
| **Campo 42** | MunicipioOcurrencia |
| Alias | Municipio Ocurrencia |
| Descripción | Describe el nombre del municipio del evento. |
| Tipo | Categórico Nominal |
| Dominio | Existen 2186 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * “Bogotá D.C” tiene la mayor cantidad de registros con 979.543 registros |
| **Campo 43** | codigoDepartamentoNotificacion |
| Alias | Código departamento notificación |
| Descripción | Describe el código del departamento del evento |
| Tipo | Categórico Nominal |
| Dominio | [Entero Positivo]  Existen 2.170 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * El código 11001 tiene la mayor cantidad de registros con 602.884 |
| **Campo 44** | DepartamentoNotificacion |
| Alias | Departamento notificación |
| Descripción | Describe el nombre del departamento del evento. |
| Tipo | Categórico Nominal |
| Dominio | Existen 2.262 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * “Bogotá D.C” tiene la mayor cantidad de registros con 1´031.909 registros |
| **Campo 45** | codigoMunicipioNotificacion |
| Alias | Código municipio notificación |
| Descripción | Describe el código del municipio del evento. |
| Tipo | Categórico Nominal |
| Dominio | [ENTERO POSITIVO]  Existen 2.264 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * “05” tiene la mayor cantidad de registros con 759.395 registros |
| **Campo 46** | MunicipioNotificacion |
| Alias | Municipio notificación |
| Descripción | Describe el nombre del municipio del evento |
| Tipo | Categórico Nominal |
| Dominio | Existen 2093 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * “Bogotá D.C” tiene la mayor cantidad de registros con 1´107.465 registros |
| **Campo 47** | codigoDepartamentoResidencia |
| Alias | Código |
| Descripción | Describe el código de residencia del propietario. |
| Tipo | Categórico Nominal |
| Dominio | [Entero Positivo]  Existen 2.092 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * “11001” tiene la mayor cantidad de registros con 678.440 registros |
| **Campo 48** | DepartamentoResidencia |
| Alias | Departamento residencia |
| Descripción | Describe el nombre del departamento de residencia |
| Tipo | Categórico Nominal |
| Dominio | Existen 2179 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * “Bogotá D.C.” tiene la mayor cantidad de registros con 1´075.205 registros |
| **Campo 49** | codigoMunicipioResidencia |
| Alias | Código |
| Descripción | Describe el código de |
| Tipo | Categórico Nominal |
| Dominio | [Entero Positivo]  Existen 2.215 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * “05” tiene la mayor cantidad de registros con 757.939 registros |
| **Campo 50** | MunicipioResidencia |
| Alias | Municipio residencia |
| Descripción | Describe el nombre del municipio de residencia |
| Tipo | Categórico Nominal |
| Dominio | Existen 2.144 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * “Bogotá D.C” tiene la mayor cantidad de registros con 1´016.830 registros |
| **Campo 51** | Área |
| Alias | Área |
| Descripción | Describe el nombre del área (Cabecera, Área Rural Dispersa, Centro Poblado) |
| Tipo | Categórico Nominal |
| Dominio | Existen 2167 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Se pueden observar varios valores atipicos, al parecer correspondiente a otras columnas. |
| **Campo 52** | CondicionFinal |
| Alias | Condición final |
| Descripción | Describe el nombre de la condición final |
| Tipo | Categórico Nominal |
| Dominio | Existen 2.259 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Se pueden observar varios valores atipicos, al parecer correspondiente a otras columnas. |
| **Campo 53** | CodigoDiagnosticoDefuncion |
| Alias | Código |
| Descripción | Describe el código de diagnóstico de defunción. |
| Tipo | Categórico Nominal |
| Dominio | [Entero Positivo]  Existen 4.601 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Se pueden observar varios valores atipicos, al parecer correspondiente a otras columnas. |
| **Campo 54** | DiagnosticoDefuncion |
| Alias | Diagnostico defunción |
| Descripción | Describe el nombre del diagnóstico de defunción |
| Tipo | Categórico Nominal |
| Dominio | Existen 4.269 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Se pueden observar varios valores atipicos, al parecer correspondiente a otras columnas. |
| **Campo 55** | TipoUPGD |
| Alias | Tipo UPGD |
| Descripción | Describe el nombre del tipo de UPGD (Unidad Primaria Generadora del Dato) |
| Tipo | Categórico Nominal |
| Dominio | Existen 4334 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Se pueden observar varios valores atipicos, al parecer correspondiente a otras columnas. |
| **Campo 56** | EstadoFinalCaso |
| Alias | Estado final Caso |
| Descripción | Describe el nombre del estado final del caso |
| Tipo | Categórico Nominal |
| Dominio | Existen 3806 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Se pueden observar varios valores atipicos, al parecer correspondiente a otras columnas. |
| **Campo 57** | Confirmado |
| Alias | Confirmado |
| Descripción | Describe el tipo de confirmación |
| Tipo | Categórico Nominal |
| Dominio | Existen 1729 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Se pueden observar varios valores atipicos, al parecer correspondiente a otras columnas. |

Tabla 4: Descripción datos fuente

### 3.3.3 Análisis preliminares

**Variables categóricas**

* Algunas columnas muestran una alta concentración de datos en categorías específicas, como el valor 1 en "PersonaBasicaID", "NO DEFINIDO" en "Prestador", "SALUD" en "Administradora", entre otras.

* Los grupos poblacionales representados en las columnas que empiezan con “GP\_” muestran que la mayoría de los casos pertenecen a la categoría "NO”, y la minoría están catalogados con "SI".
* Las columnas de fechas muestran una distribución desigual, con muchas entradas en 1900-01-01, que es un valor predeterminado que puede indicar una fecha desconocida o no especificada. Las fechas más recientes también son frecuentes.

**Comentarios adicionales**

* No se incluye en el análisis la columna “\_\_null\_dask\_index\_\_” puesto que solo es un id único para cada registro.
* No se cargaron los datos del 2023.
* Hay múltiples inconsistencias de los datos iniciales.
* Hay un gran número de registros con valores por defecto, en especial las relacionadas a códigos y fechas.

## 3.4 Análisis de correlación

Para esta fuente de datos no se realizó una exploración de la relación entre pares de variables numéricas mediante el cálculo de matriz de correlación, porque la mayoría de las variables fueron catalogadas como categóricas.