Imagen que contiene Forma

Descripción generada automáticamente

**PROYECTO AGORA**

**PREPARACIÓN DE LOS DATOS**

**Agosto 2024**

**Control de cambios**

|  |  |  |  |
| --- | --- | --- | --- |
| **VERSIÓN** | **FECHA** | **MODIFICACIÓN** | **RESPONSABLE** |
| 1.0 | 24 de junio de 2024 | Modificación del documento | Luis Carlos Chicaíza |
| 1.1 | 26 de junio de 2024 | Se documenta las recomendaciones de RIPS | Luis Carlos Chicaíza |
| 1.1. | 28 de junio de 2024 | Se revisa y ajusta capítulo 1 y 2 | Pedro Fabián Pérez Arteaga |
| 1.1. | 1 de julio de 2024 | Se documenta las recomendaciones de ESTADÍSTICAS VITALES | Luis Carlos Chicaíza |
| 1.2 | 4 de julio de 2024 | Se documenta nuevas fuentes de datos | Luis Carlos Chicaíza |
| 1.2 | 30 de julio de 2024 | Se revisa y hacen ajustes finales | Pedro Fabián Pérez Arteaga |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Contenido

[1 Introducción 6](#_Toc171274531)

[2 Metodología 7](#_Toc171274532)

[2.1 Evaluación de los datos 7](#_Toc171274533)

[3 Listado de fuentes 8](#_Toc171274534)

[3.1 RIPS 8](#_Toc171274535)

[3.2 ESTADÍSTICAS VITALES 8](#_Toc171274536)

[3.3 SEGCOVID 9](#_Toc171274537)

[3.4 SIVIGILA 9](#_Toc171274538)

[3.5 VACUNAS COVID 10](#_Toc171274539)

[3.6 Actualización de fuentes 10](#_Toc171274540)

[4 Datos proyecto AGORA 11](#_Toc171274541)

[4.1 RIPS 11](#_Toc171274542)

[4.1.1 Ajustes iniciales 12](#_Toc171274543)

[4.1.2 Análisis preliminar 13](#_Toc171274544)

[4.1.3 Preparación de datos 14](#_Toc171274545)

[4.2 ESTADÍSTICAS VITALES 16](#_Toc171274546)

[4.2.1 Ajustes iniciales 16](#_Toc171274547)

[4.2.2 Análisis preliminar 16](#_Toc171274548)

[4.2.3 Preparación de datos 18](#_Toc171274549)

[4.3 SEGCOVID 19](#_Toc171274550)

[4.3.1 Ajustes iniciales 19](#_Toc171274551)

[4.3.2 Análisis preliminar 19](#_Toc171274552)

[4.3.3 Preparación de datos 21](#_Toc171274553)

[4.4 SIVIGILA 23](#_Toc171274554)

[4.4.1 Ajustes iniciales 23](#_Toc171274555)

[4.4.2 Análisis preliminar 23](#_Toc171274556)

[4.4.3 Preparación de datos 25](#_Toc171274557)

[4.5 VACUNAS COVID 27](#_Toc171274558)

[4.5.1 Ajustes iniciales 27](#_Toc171274559)

[4.5.2 Análisis preliminar 27](#_Toc171274560)

[4.5.3 Preparación de datos 28](#_Toc171274561)

[5 Conclusiones 29](#_Toc171274562)

**Ilustraciones**

[Ilustración 1 Muestra de 5 registros duplicados exactamente 13](#_Toc171274563)

[Ilustración 2 Cálculo de registros duplicados por año 13](#_Toc171274564)

[Ilustración 3 Valores no numéricos 16](#_Toc171274565)

[Ilustración 4 Datos inconsistentes variable Peso 17](#_Toc171274566)

[Ilustración 5 Valores no numéricos SEGCOVID 20](#_Toc171274567)

[Ilustración 6 Datos nulos 21](#_Toc171274568)

[Ilustración 7 Valores no numéricos variable Edad 24](#_Toc171274569)

[Ilustración 8 Nuevos datos variable Edad 24](#_Toc171274570)

[Ilustración 9 Nombres de las fuentes VACUNAS COVID 27](#_Toc171274571)

**Tablas**

[Tabla 1 Descripción inicial de la fuente RIPS 8](#_Toc171274572)

[Tabla 2. Descripción inicial de la fuente ESTADÍSTICAS VITALES 8](#_Toc171274573)

[Tabla 3. Descripción inicial de la fuente SEGCOVID 9](#_Toc171274574)

[Tabla 4. Descripción inicial de la fuente SIVIGILA 9](#_Toc171274575)

[Tabla 5. Descripción inicial de la fuente VACUNASCOVID 10](#_Toc171274576)

[Tabla 6. Verificación de diferentes nombres de columnas fuente RIPS 12](#_Toc171274577)

# Introducción

Este documento muestra los principales hallazgos derivados del proceso de exploración inicial de los datos. Este proyecto se basa en seis bases de datos clave: RIPS, ESTADÍSTICAS VITALES (Nacimientos y Defunciones), SEGCOVID, SIVIGILA y VACUNASCOVID, configuradas con diferentes tipos de información crucial, definidas en sus columnas. La fase de exploración de los datos fue primordial para el análisis de la completitud y la calidad de los datos.

Durante esta fase inicial de exploración, se identificó la fuente de datos y se examinaron las variables contenidas en ella junto con sus características principales. Se realizó un análisis exhaustivo de las variables disponibles, aplicando técnicas estadísticas descriptivas básicas para evaluar la distribución, la variabilidad y las tendencias presentes en los registros. Este análisis preliminar permitió entender la naturaleza de los datos y permitió identificar posibles áreas de interés y enfoque para análisis posteriores, y los cambios o modificaciones a estas fuentes de datos para consolidar la información, eliminando datos incongruentes o incompletos.

En las siguientes secciones, se detallan tres secciones principales. Primero los resultados de la metodología la exploración de los datos destacando los principales cambios necesarios para limpiar y preparar la información. Estos hallazgos iniciales sientan las bases para el desarrollo de recomendaciones y estrategias de preparación de datos aplicables a todas las bases de datos incluidas en este proyecto. Segundo documentar la selección por parte de los investigadores del proyecto AGORA de las variables no relevantes que deben filtrarse en cada fuente de datos y la creación de variables derivadas necesarias en cada una de las fuentes. Tercero el desarrollo de los ETL's que transforman los datos de rawdata a stagedata para ser los datos limpios del proyecto para la fase de visualización de datos, que incluyen los ajustes acordados en la primera y segunda parte.

# Metodología

Según la exploración de los datos provenientes de cada fuente de información, han permitido analizar la calidad e integridad de los datos proporcionados. En el capítulo 3 se destacarán los resultados más significativos de la fase de exploración y se definirán las recomendaciones específicas para el tratamiento y limpieza de los datos. Si la información se utilizará para aplicar algoritmos de predicción o analítica avanzada, se debe realizar una preparación exhaustiva, teniendo en cuenta las precondiciones necesarias para el funcionamiento óptimo de estos algoritmos sobre la fuente inicial de datos.

## Evaluación de los datos

En la fase de exploración, se evaluaron varios aspectos cruciales para asegurar la calidad y utilidad de los datos. Primero, se verificó la completitud de los datos, asegurándose de que no hubiera valores faltantes significativos que pudieran afectar el análisis. En segundo lugar, se revisó la precisión de los datos para garantizar que reflejaran correctamente la realidad y no contuvieran errores que pudieran distorsionar los resultados. Además, se comprobó la consistencia de los datos, revisando que fueran coherentes a través de todas las fuentes y registros. Finalmente, se confirmó la actualidad de los datos, verificando que estuvieran actualizados y fueran relevantes para el análisis en curso. Estos aspectos fueron fundamentales para garantizar la integridad y la fiabilidad del análisis de datos.

# Listado de fuentes

Teniendo en cuenta las fuentes de datos trabajadas, a continuación, se menciona un cuadro detallado por cada una de las fuentes de datos:

## RIPS

|  |  |
| --- | --- |
| **FUENTE** | **INFORMACIÓN RIPS** |
| Descripción | Contiene los datos del sistema RIPS (Registro Individual de Prestación de Servicios de Salud) |
| Período | 2009 - 2022 |
| Extensión | .parquet |
| Volumen (#filas) | 4´262.537.750 (después de limpieza básica) |
| Campos (#columnas) | 21 |
| Fecha recepción | 7 de diciembre de 2023 |

Tabla 1 Descripción inicial de la fuente RIPS

## ESTADÍSTICAS VITALES

|  |  |
| --- | --- |
| **FUENTE** | **INFORMACIÓN ESTADÍSTICAS VITALES** |
| Descripción | Contiene los datos del sistema ESTADÍSTICAS VITALES |
| Período | 2014 - 2023 |
| Extensión | .parquet |
| Volumen (#filas) | 5.458.140 |
| Campos (#columnas) | 12 |
| Fecha recepción | 7 de diciembre de 2023 |

Tabla 2. Descripción inicial de la fuente ESTADÍSTICAS VITALES

## SEGCOVID

|  |  |
| --- | --- |
| **FUENTE** | **INFORMACIÓN SEGCOVID** |
| Descripción | Contiene los datos obtenidos del seguimiento de los casos de COVID-19 denominado como SEGCOVID |
| Período | 2014 – 2023 |
| Extensión | .parquet |
| Volumen (#filas) | 18’431.733 |
| Campos (#columnas) | 40 |
| Fecha recepción | 7 de diciembre de 2023 |

Tabla 3. Descripción inicial de la fuente SEGCOVID

## SIVIGILA

|  |  |
| --- | --- |
| **FUENTE** | **INFORMACIÓN SIVIGILA** |
| Descripción | Contiene los datos del sistema SIVIGILA |
| Período | 2008 - 2022 |
| Extensión | .parquet |
| Volumen (#filas) | 8’836.699 |
| Campos (#columnas) | 57 |
| Fecha recepción | 7 de diciembre de 2023 |

Tabla 4. Descripción inicial de la fuente SIVIGILA

## VACUNAS COVID

|  |  |
| --- | --- |
| **FUENTE** | **INFORMACIÓN VACUNASCOVID** |
| Descripción | Contiene los datos del sistema VACUNASCOVID |
| Período | 2020 - 2023 |
| Extensión | .parquet |
| Volumen (#filas) | 13.239.573 |
| Campos (#columnas) | 64 |
| Fecha recepción | 7 de diciembre de 2023 |

Tabla 5. Descripción inicial de la fuente VACUNASCOVID

## Actualización de fuentes

Se recibe una actualización de dos fuentes, Vitales (datos de defunciones) y Vacunas COVID versión 2.

# Datos proyecto AGORA

## RIPS

La fuente de datos denominada RIPS (Registro Individual de Prestación de Servicios de Salud), comprende información recopilada durante el periodo comprendido entre 2009 y 2022. Este periodo permitió analizar una variedad de comportamientos para la comprensión y el análisis de las tendencias relacionadas con la prestación de servicios de salud en Colombia.

Durante la fase inicial de exploración, se realizó un análisis exhaustivo de las variables contenidas en el conjunto de datos del sistema RIPS. 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.

### Ajustes iniciales

El total de registros es de 4’262.860.903 registros con 21 variables. Los datos se comprenden desde el 2009 hasta el 2022.

En este periodo, se observa una variación en los nombres de las variables registradas en el sistema RIPS. Esta variabilidad en los nombres de las variables se refleja en la siguiente tabla:

|  | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CausaExternaCD | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí |
| Codigo | Sí | Sí | Sí | Sí | Sí | Sí | No | No | No | No | No | No | No | No |
| codigo | No | No | No | No | No | No | Sí | Sí | Sí | Sí | No | No | No | No |
| CodigoAdm | No | No | No | No | No | No | No | No | No | No | Sí | Sí | Sí | Sí |
| CodigoProcedimiento | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí |
| CostoConsulta | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí |
| CostoProcedimiento | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí |
| DxEgreso | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí |
| DxPrincipal | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí |
| EstadoSalidaCD | Sí | Sí | Sí | Sí | Sí | Sí | No | No | No | No | No | No | No | No |
| estadosalidaCD | No | No | No | No | No | No | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí |
| FinalidadConsultaCD | Sí | Sí | Sí | Sí | Sí | Sí | No | No | No | No | Sí | Sí | Sí | Sí |
| FinalidadConsultacd | No | No | No | No | No | No | Sí | Sí | Sí | Sí | No | No | No | No |
| FinalidadProcedimientoCD | No | No | No | No | No | No | Sí | Sí | Sí | Sí | No | No | No | No |
| FinalidadProcedimientosCD | Sí | Sí | Sí | Sí | Sí | Sí | No | No | No | No | Sí | Sí | Sí | Sí |
| MunicipioCD | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí |
| NumeroDiasEstancia | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí |
| PersonaID | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí |
| Prestador | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí |
| SexoDesc | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | No |
| Sexodesc | No | No | No | No | No | No | No | No | No | No | No | No | No | Sí |
| TipoDiagnosticoPrincipalCD | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí |
| TipoEventoRIPSDesc | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí |
| TipoUsuarioCD | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí |
| edad | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí |
| fechaid | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí | Sí |
| NetoAPagarConsulta | No | No | No | No | No | No | No | No | No | No | Sí | Sí | Sí | Sí |

Tabla 6. Verificación de diferentes nombres de columnas fuente RIPS

Se realizó una unificación de las variables y se adiciona la variable *NetoaPagarConsulta* a los años 2009 al 2018 para el análisis descriptivo de la fuente.

### Análisis preliminar

**Registros duplicados**

Es importante validar la fuente de datos de registros exactamente iguales o duplicados de todas las columnas. Se evidencian registros duplicados, por lo cual la mejor opción es que deberían ser eliminados y dejar solamente un registro, aunque si los datos son para realizar análisis analíticos robustos, no deberían influir significativamente en el resultado final.

Texto

Descripción generada automáticamente

Ilustración 1 Muestra de 5 registros duplicados exactamente

Procesando los datos por cada año y validando el número de duplicados se obtiene un total de 323.129 registros. En el análisis exploratorio de los datos se eliminaron los registros duplicados.

Texto

Descripción generada automáticamente con confianza media  
Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Ilustración 2 Cálculo de registros duplicados por año

### Preparación de datos

En general, la fuente de datos evidencia problemas de inconsistencia de los datos, se mencionan algunas observaciones para considerar por un experto.

Variables numéricas

**Eliminación de Valores Negativos**

En la variable “*Edad”*, se deben eliminar o corregir los valores negativos ya que no es posible que una persona tenga una edad negativa.

En las variables relacionadas a “*Costos*” (costoconsulta, costoprocedimiento, netoapagarconsulta), es necesario eliminar o corregir los valores negativos, puesto que los costos no pueden ser valores negativos.

En la variable "*numerodiasestancia*". Se debe verificar y eliminar también los valores negativos.

**Corrección de Valores Extremos**

En la variable “Edad” se deben revisar y corregir los valores de edad superiores a 120 años, ya que se identificó ciertos valores atípicos considerados como razonables para la edad humana.

En los costos y neto a pagar existen valores extremadamente altos en las variables de costos por lo cual deberían ser investigados. Podrían ser errores de entrada de datos y necesitar corrección.

**Manejo de Valores Faltantes**

En la variable “netoapagarconsulta” tiene menos datos que las otras, esto se debe a que esta variable se empezó a diligenciar desde el 2019. Se pueden imputar los valores faltantes utilizando la mediana, pero probablemente la mejor opción sea eliminar las filas y columnas con muchos valores faltantes.

Variables Categóricas

**Estandarización de los valores**

En la columna “sexodesc” los valores deben ser homogéneos. Es decir, se recomienda estandarizar los valores como "FEMENINO" y "MASCULINO". Los valores como "F", "M", "N", "NO DEFINIDO", "INDEFINIDO", e "INDETERMINADO" deben ser corregidos o recategorizados.

**Eliminación o corrección de Valores Nulos**

Se evidencio una gran cantidad de variables con valores NULL, se deberia investigar por qué existen estos valores y determinar si deben ser imputados o eliminados. Las columnas identificadas que requieren una especial atención son:

* tipoeventoripsdesc
* finalidadprocedimientoscd
* finalidadconsultacd
* tipousuariocd
* causaexternacd
* estadosalidacd
* tipodiagnosticoprincipalcd
* municipiocd
* sexodesc

**Valores frecuentes atípicos**

La columna “personaid” se observa un valor 1 que podría entenderse que es un valor predeterminado, por lo que muchos registros tienen este valor. Se recomienda verificar y posiblemente eliminar o corregir estas entradas.

**Revisión de las variables relacionada a Códigos**

En la columna “código” Se debería verificar los códigos y asegurarse de que no existan duplicados que representen la misma entidad.

En las columnas “dxprincipal” y “dxegreso” se debería verificar y corregir cualquier código que sea claramente un error (por ejemplo, un valor de 1 en dxprincipal mencionado anteriormente).

En la columna “municipiocd” se debería corregir los valores NULL y asegurarse de que los códigos sean válidos.

**Manejo de Valores NULL**

Se identificaron columnas con valores NULL significativos, en algunos casos el porcentaje de valores NULL es muy alto, como en “estadosalidacd” y “tipodiagnosticoprincipalcd”, se debe decidir si estas columnas son realmente necesarias o si pueden ser eliminadas.

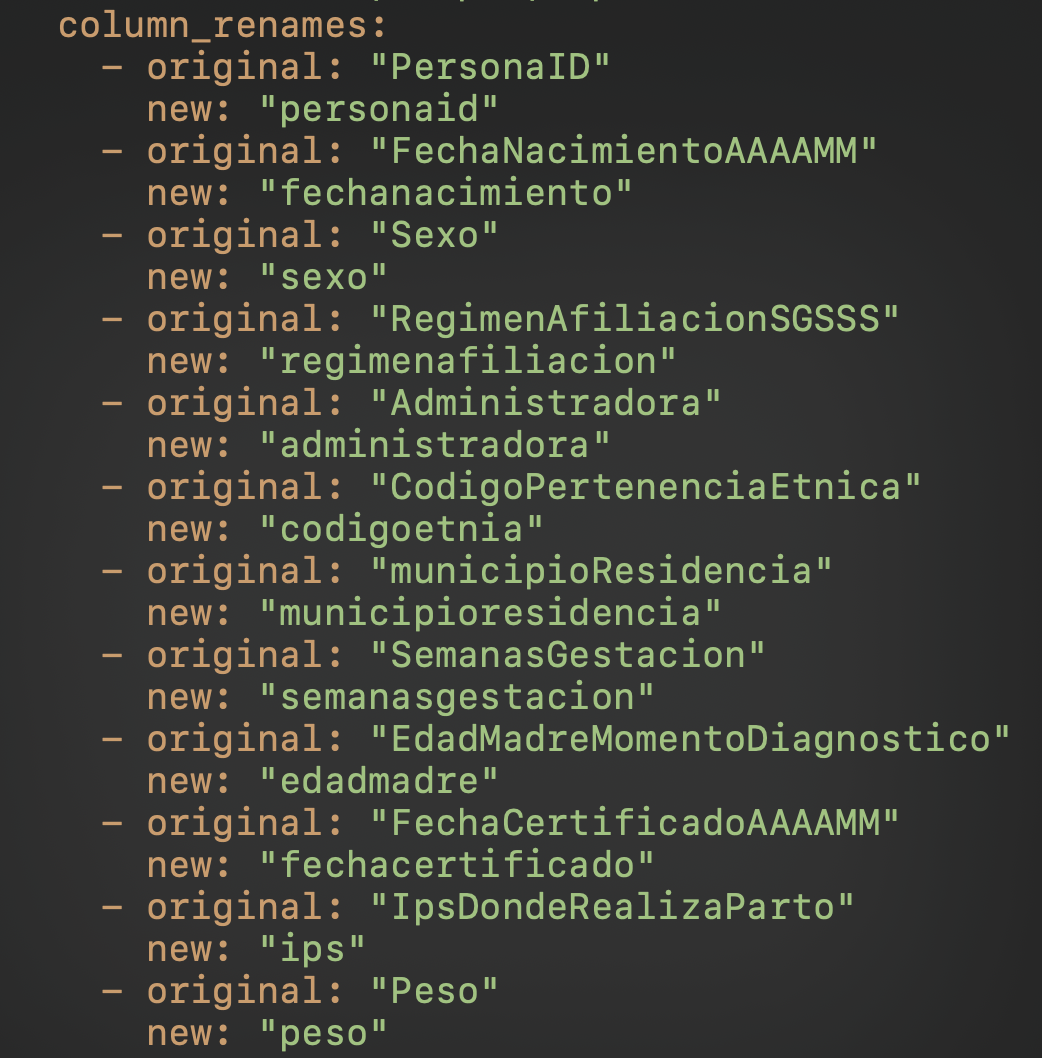
## ESTADÍSTICAS VITALES

La fuente de datos denominada ESTADÍSTICAS VITALES, comprende información recopilada durante el periodo comprendido entre 2014 a 2023. Este periodo permitió analizar una variedad de comportamientos para la comprensión y el análisis de tendencias.

Durante la fase inicial de exploración, se realizó un análisis exhaustivo de las variables contenidas en el conjunto de datos del sistema ESTADÍSTICAS VITALES. 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.

### Ajustes iniciales

La información recibida abarca un periodo que comprende desde 2014 hasta 2023. En el proceso de exploración de datos, se realiza un ajuste de nombre a minúsculas, se realiza acotación de nombres de las variables. Finalmente la información recibida no evidencia registros iguales o duplicados.



### Análisis preliminar

**Ajustes requeridos en registros**

Al analizar los valores no numéricos de los datos originales de ESTADÍSTICAS VITALES se identificó que en las variables 'SemanasGestacion' y ´FechaCertificadoAAAAMM´ hay 2 datos erróneos, mientras que en la variable 'Peso' hay 64.630, por lo cual estas variables deberían ser catalogadas como variables numéricas.

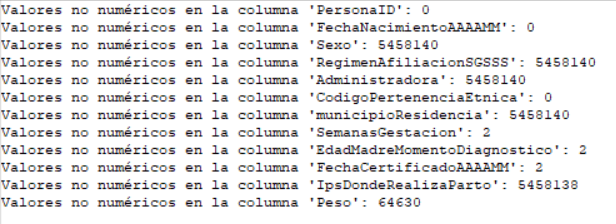


Ilustración 3 Valores no numéricos

Al comprobar estos registros se validó que la información tenía estos datos atípicos, que iban a afectar directamente en el cálculo de las estadísticas básicas

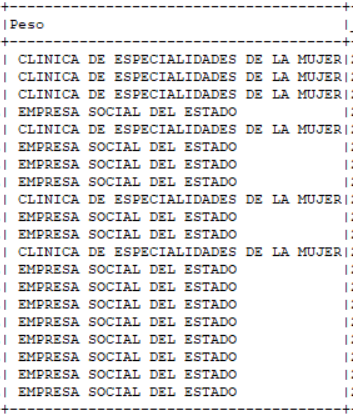


Ilustración 4 Datos inconsistentes variable Peso

En el proceso de exploración de datos, en total se eliminaron 64.632 registros de la muestra original, que se componen de 64.630 registros de Peso y 2 registros que compartían el mismo error en las columnas ‘SemanasGestacion' y 'EdadMadreMomentoDiagnostico', por lo cual la fuente final quedo con un total de 5.393.508 registros. Estos registros son importantes considerar en el proceso de preparación de datos.

### Preparación de datos

En general, la fuente de datos evidencia problemas de inconsistencia de los datos, se mencionan algunas observaciones para considerar por un experto.

Variables numéricas

**Corrección de Valores Extremos**

En la variable “SemanasGestacion” los valores de semanas de gestación deberían estar entre 22 y 42 semanas para ser considerados normales, en algunos casos se evidencia valores por fuera de ese rango.

En la variable “EdadMadreMomentoDiagnostico” existen valores que podrían ser atípicos, en algunos casos se evidencia valores de madres con 11 años o mayores a 54 años.

En la variable “Peso” existen valores que podrían ser atípicos, en algunos casos se evidencia valores de pesos extremadamente bajos y también valores extremadamente altos. Se evidenciaron pesos de 675g y 4850 gramos.

**Manejo de Valores Faltantes**

Se pueden imputar los valores faltantes utilizando la mediana, pero probablemente la mejor opción sea eliminar las filas y columnas con muchos valores faltantes.

Se puede comprobar la consistencia de los datos entre diferentes columnas relacionadas, por ejemplo, la edad de la madre y semanas de gestación deberían ser coherentes.

En general, identificar los datos inconsistentes según un experto, con el propósito de corregirlos o eliminarlos según sea el caso.

Variables Categóricas

**Eliminación o corrección de valores nulos**

No se evidenciaron registros con valores en nulo después de los ajustes iniciales.

Sin embargo, sin los ajustes iniciales se debe considerar revisar la columna “Peso” por los posibles datos no numéricos que se identificaron al inicio, también para las columnas 'SemanasGestacion' y ´FechaCertificadoAAAAMM’ por los registros atípicos identificados.

## SEGCOVID

La fuente de datos denominada SEGCOVID, comprende información recopilada durante el periodo comprendido entre 2020 a 2023. Este periodo permitió analizar una variedad de comportamientos para la comprensión y el análisis de tendencias.

Durante la fase inicial de exploración, se realizó un análisis exhaustivo de las variables contenidas en el conjunto de datos, cabe mencionar que la información de SEGCOVID se generó a partir de una aplicación en donde se registraba el seguimiento a los casos de COVID-19. 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.

### Ajustes iniciales

La información recibida abarca un periodo que comprende desde 2020 hasta 2023. En el proceso de exploración de los datos, 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. El total de registros es de 18’431.733 registros en el seguimiento de los casos de COVID-19 conocidos como SEGCOVID.

### Análisis preliminar

**Ajustes requeridos en registros**

Al realizar el proceso exploratorio de los datos en SEGCOVID se identificó que todas las columnas corresponden a valores categóricos a excepción de 'PersonaBasicaID', es decir, es decir que no hay otras variables con valores numéricos, se puede evidenciar que todas las variables tienen el mismo número de registros y son catalogadas como categóricas.

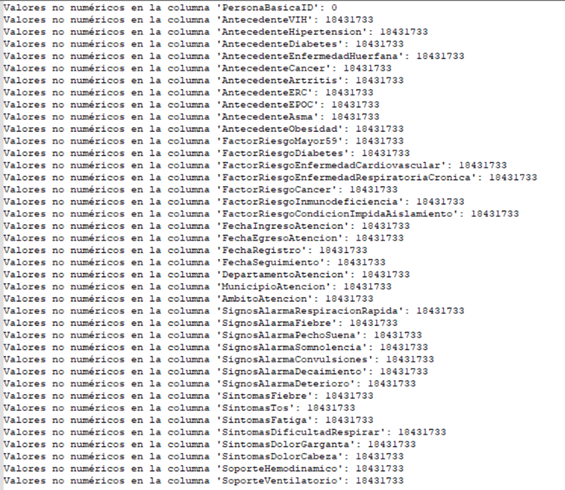


Ilustración 5 Valores no numéricos SEGCOVID

Al comprobar estos registros también se validó si había datos en nulo, de lo cual no se evidencio algún registro con esta característica.

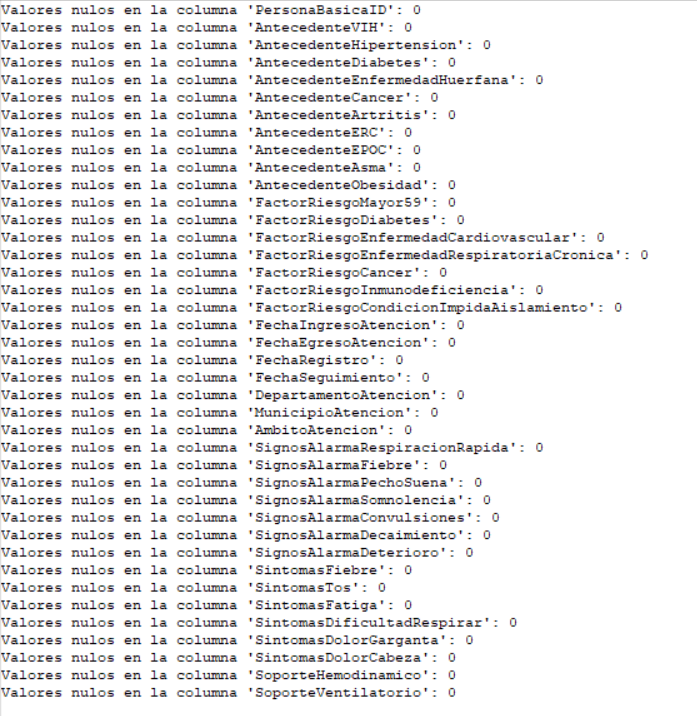


Ilustración 6 Datos nulos

Por lo tanto, no fue necesario modificar la fuente original en el proceso exploratorio de los datos y este se realizó con los 18’431.733 registros disponibles.

### Preparación de datos

En general la fuente de datos no evidencia problemas de inconsistencia de los datos, sin embargo, se mencionan algunas observaciones para tener en cuenta por un experto:

**Verificación de valores**

Se evidencio una gran cantidad de registros con valores 'No', se destacan los síntomas,factores y antecedentes, se podría investigar si es correcto que en su gran mayoría tengan este valor. Las columnas identificadas son las siguientes:

* SignosAlarmaRespiracionRapida
* SignosAlarmaFiebre, SignosAlarmaPechoSuena,
* SignosAlarmaSomnolencia
* SignosAlarmaConvulsiones
* SignosAlarmaDecaimiento
* SignosAlarmaDeterioro
* SintomasFiebre
* SintomasTos
* SintomasFatiga
* SintomasDificultadRespirar
* SintomasDolorGarganta
* SintomasDolorCab
* AntecedenteVIH
* AntecedenteHipertension
* AntecedenteDiabetes
* AntecedenteEnfermedadHuerfana
* AntecedenteCancer
* AntecedenteArtritis
* AntecedenteERC
* AntecedenteEPOC
* AntecedenteAsma
* AntecedenteObesidad
* FactorRiesgoMayor59
* FactorRiesgoDiabetes
* FactorRiesgoEnfermedadCardiovascular
* FactorRiesgoEnfermedadRespiratoriaCronica
* FactorRiesgoCancer
* FactorRiesgoInmunodeficiencia
* FactorRiesgoCondicionImpidaAislamiento

Las variables FechaIngresoAtencion, FechaEgresoAtencion, FechaRegistro y FechaSeguimiento, se evidencio que hay un número significativo de fechas marcadas como '1900-01-01', que pueden indicar un valor predeterminado o este valor pudo ser asignado como dato faltante.

En general, no se evidenciaron registros con valores en nulo o duplicados en esta fuente de datos.

## SIVIGILA

La fuente de datos denominada SIVIGILA, comprende información recopilada durante el periodo comprendido entre 2008 a 2022. Este periodo permitió analizar una variedad de comportamientos para la comprensión y el análisis de tendencias de la fuente de datos.

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.

### Ajustes iniciales

La información recibida abarca un periodo que comprende desde 2008 hasta 2022. En el proceso de exploración de datos 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 columnas.

### Análisis preliminar

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

Ilustración 7 Valores no numéricos variable Edad

Por lo que se modificó esta columna para extraer los valores numéricos correspondientes a la edad y así generar el análisis estadístico correspondiente.

Tabla

Descripción generada automáticamente

Ilustración 8 Nuevos datos variable Edad

### Preparación de datos

En general la fuente de datos evidencia problemas de inconsistencia de los datos, se mencionan algunas observaciones para tener en cuenta por un experto:

Variables numéricas

**Corrección de Valores Extremos**

Se identificaron 36597 valores en nulo en la variable edad, por lo cual se deben eliminar estos registros o imputar estos valores.

En la variable “Edad”, los valores de edad deberían estar en un rango definido por un experto (por ejemplo, de 0 a 120 años). Sin embargo, en la exploración se identificó un registro con una edad de 95000, lo cual es claramente un error y valores se deben tratar. Se recomienda identificar y corregir estos valores atípicos.

En la variable “SemanaEpidemiologica”, los valores deberían estar en un rango definido por un experto (por ejemplo, de 1 a 53 semanas). Es importante verificar que todos los valores se encuentran dentro del rango definido y así poder corregir cualquier dato que no cumpla con este criterio. Esto podría eliminar registros con estos valores atípicos o imputarlos con un valor valido.

**Manejo de Valores Faltantes**

En la variable “Edad”, los valores faltantes se pueden imputar utilizando la mediana o la media de la edad, asegurando que la imputación se mantenga dentro del rango definido por el experto.

En la variable “SemanaEpidemiologica”, los valores faltantes pueden ser imputados utilizando la mediana de las semanas epidemiológicas reportadas.

**Transformaciones Necesarias**

Para la variable “Edad”, después de corregir el formato de la edad puesto que está definido con una palabra según lo observado en el análisis preliminar, y de identificar los valores atípicos y manejar los valores faltantes, se recomienda normalizar o estandarizar la variable para mejorar la consistencia.

En la variable “SemanaEpidemiologica”, es importante asegurar que los valores estén en un formato consistente, preferiblemente como enteros, para facilitar su uso en análisis temporales y epidemiológicos.

**Validación y Consistencia**

Después de realizar las correcciones y transformaciones necesarias, se debe validar que los datos estén en el formato esperado y sean coherentes con otros datos relacionados. En general, es crucial contar con expertos en el dominio para identificar y corregir datos inconsistentes o erróneos, asegurando así la calidad del conjunto de datos final.

Variables Categóricas

**Corrección de fuente de datos**

Se evidencio que la fuente original tiene algunos datos inconsistentes en sus registros, en el caso de "Bogotá, D.C" el delimitador con ', ' se utiliza para cargar los datos correspondientes de cada columna, en este caso hace que estos registros se corran los valores de las columnas. Gran cantidad de registros se ven afectados por este problema.

Por ejemplo, en 'CodigoDepartamentoOcurrencia' y 'CodigoMunicipioOcurrencia' se debe verificar la integridad y validez de los códigos de departamento y municipio. Se debe evaluar con un experto, si no deberían encontrarse registros como "NO" o "SI" que puedan estar contaminando los datos.

**Manejo de Valores Faltantes**

Se identificaron registros con valores NULL, se debe evaluar y considerar si deben ser imputados, eliminados o si se puede obtener esta información de otras fuentes, algunas variables afectadas son las siguientes:

* EstadoFinalCaso
* Confirmado,
* CodigoPaisOcurrencia
* CodigoDepartamentoOcurrencia
* CodigoMunicipioOcurrencia
* CodigoDepartamentoNotificacion
* CodigoMunicipioNotificacion
* CodigoDepartamentoResidencia
* CodigoMunicipioResidencia
* CodigoDiagnosticoDefuncion

## VACUNAS COVID

Para el procesamiento de los datos, se definió unos contendores en la infraestructura con la información por año. Esta información está consolidada en su mayoría desde el año 2020 hasta el año 2023. Existen otros registros fuera de este rango de fechas 2024-2030 (36 registros), por lo que no se consideraron en la fase de exploración.

### Ajustes iniciales

El total de registros de toda la fuente es de 13´239.609 registros y 64 variables. En la fase inicial de exploración se utilizaron los datos de los años 2020 a 2023, del cual se realizó un análisis exploratorio con 13´239.573 registros. En el análisis exploratorio se omitieron 5 registros del año 2019 que están en su mayoría en valores NULL, los cuales afectaban en las estadísticas de algunas variables.

### Análisis preliminar

**Ajustes requeridos en nombres de las fuentes**

En el proceso de exploración la información contempló el periodo comprendido desde el 2020 hasta 2023. No fue necesario aplicar ajuste en los nombres de las variables y no se evidencio un numero o nombre distinto de las columnas a través del tiempo. En general, se identificaron 64 variables en la fuente VACUNAS COVID.

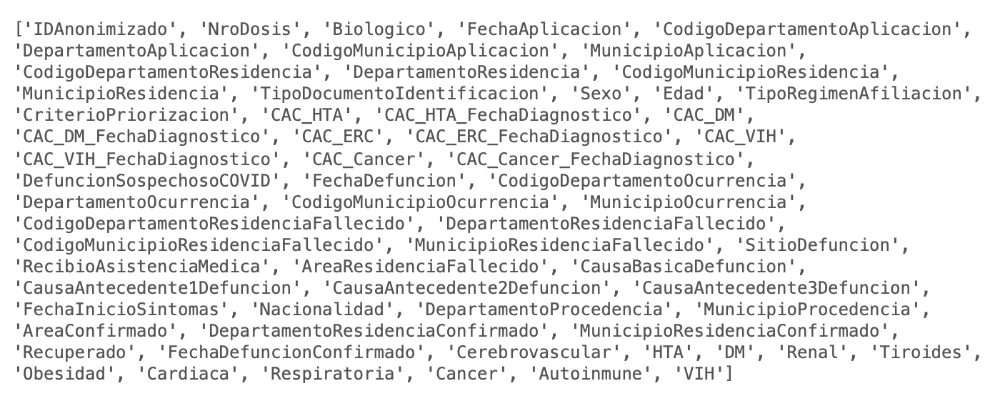


Ilustración 9 Nombres de las fuentes VACUNAS COVID

En el proceso de exploración no se evidencio registros duplicados por año.

### Preparación de datos

En general la fuente de datos evidencia problemas de inconsistencia de los datos, se mencionan algunas observaciones para tener en cuenta por un experto:

Variables numéricas

En la variable ‘Edad’ existen registros nulos o no definidos, por lo que se deberían eliminar los registros o imputar.

Variables Categóricas

**Estandarización de Valores**

En la columna "CESAR", los valores NULL deben ser evaluados para imputación o exclusión según la relevancia de la variable en el análisis. Es recomendable estandarizar los valores restantes para consistencia en análisis posteriores.

**Eliminación o Corrección de Valores Nulos**

En variables como "SitioDefuncion", se puede considerar la imputación basada en la información disponible en otros campos relacionados, como por ejemplo el área de residencia o el departamento de residencia.

En el caso de "CausaAntecedente3Defuncion", se puede analizar con un experto la posibilidad de eliminar estos registros si la falta de datos afecta significativamente al análisis o también considerar métodos avanzados de imputación si los datos son críticos para análisis posteriores.

**Manejo de Valores NULL**

Por ejemplo, variables como " SitioDefuncion" y "SEXO" muestran una cantidad considerable de valores nulos que requieren evaluación. Se debe decidir si estas columnas son esenciales para el análisis o pueden ser eliminadas tras una revisión exhaustiva.

En general, hay muchos valores en nulo en la fuente de datos que deberían tratarse. La información analizada indica que hay un problema de calidad en los datos recopilados, con registros incompletos o incluso inexactos.

# Conclusiones

Durante el proceso de análisis y preparación de datos de las fuentes proporcionadas al equipo de ALIANZA CAOBA, se identificaron varias inconsistencias comunes que afectaron la integridad y la consistencia de las fuentes de datos suministradas. Se observaron problemas significativos como valores nulos en múltiples columnas clave, así como valores atípicos en variables numéricas que requieren corrección. Por ejemplo, se encontraron registros con edades improbables o inconsistencias en las fechas de registro. Además, la estandarización de los valores categóricos, como género y códigos, mostró algunas discrepancias que se deberían abordarse para asegurar la coherencia en los análisis posteriores y así poder generar resultados exactos y consistentes.

Es fundamental que un experto que conozca estos datos pueda seguir estas recomendaciones y así implementar estrategias adecuadas para manejar los datos faltantes, ya sea mediante imputación o eliminación selectiva de registros. Asimismo, se recomienda una revisión exhaustiva de las variables relacionadas con códigos y fechas, para la mayoría de las fuentes no se evidencio un número significativo de registros duplicados, sin embargo, es importante realizar un análisis mucho más exhaustivo para validar que no existan duplicados ni valores fuera de contexto. Este enfoque desarrollado desde la fase exploratoria de los datos hasta las respectivas recomendaciones no solo mejorará la calidad de los datos, sino que también fortalecerá la validez y confiabilidad de cualquier análisis futuro realizado con estas fuentes de información crítica para la salud pública y otras áreas sensibles.

Estas recomendaciones ayudarán a mejorar la calidad de los datos y asegurarán que futuros análisis sean más precisos y significativos.