Imagen que contiene Forma

Descripción generada automáticamente

**PROYECTO AGORA**

**ENTENDIMIENTO DE LOS DATOS - RIPS**

**DOCUMENTO INICIAL**

**2024**

**Control de cambios**

|  |  |  |  |
| --- | --- | --- | --- |
| **VERSIÓN** | **FECHA** | **MODIFICACIÓN** | **RESPONSABLE** |
| 1.0 | 20 de marzo de 2024 | Documento inicial | Luis Carlos Chicaíza |
| 1.1 | 22 de marzo de 2024 | Se documentan características de los datos | Luis Carlos Chicaíza |
| 1.2 | 26 de marzo de 2024 | Se documentan características de los datos | Luis Carlos Chicaíza |
| 1.2.1 | 04 de abril de 2024 | Se documenta matriz de correlación | Luis Carlos Chicaíza |
| 1.2.2 | 09 de abril de 2024 | Revisión y ajustes y decisión de cambios en estructura | Pedro Pérez y Johan Calderón |
| 1.3 | 24 de abril de 2024 | Modificación del documento | Luis Carlos Chicaíza |
| 1.4 | 25 de abril de 2024 | Ajustes complementarios finales y ampliación de interpretación de gráficas | Pedro Pérez |
| 1.5 | 07 de mayo de 2025 | Ajustes del documento con base a los comentarios recibidos | Luis Carlos Chicaíza |

Contenido

[1 Introducción 4](#_Toc164932060)

[2 Metodología 5](#_Toc164932061)

[2.1 Exploración de los datos 5](#_Toc164932062)

[3 Fuente principal de los datos 6](#_Toc164932063)

[3.1 Origen y procedencia de los datos 6](#_Toc164932064)

[3.2 Descripción general de RIPS y ajustes iniciales 7](#_Toc164932065)

[3.2.1 Ajustes requeridos en nombres de las fuentes 8](#_Toc164932066)

[3.2.2 Ajustes requeridos en registros duplicados 9](#_Toc164932067)

[3.2.3 Ajustes requeridos en registros de resumen en la exportación 10](#_Toc164932068)

[3.3 Estadísticas descriptivas de las variables 10](#_Toc164932069)

[3.3.1 Esquema para la exploración y análisis descriptivo de los datos 11](#_Toc164932070)

[3.3.2 Listado detallado de la fuente 12](#_Toc164932071)

[3.3.3 Análisis preliminares 27](#_Toc164932072)

[3.4 Análisis de correlación 28](#_Toc164932073)

[3.4.1 ¿Qué es el Coeficiente de Correlación? 28](#_Toc164932074)

[3.5 Visualización de datos 29](#_Toc164932075)

[3.5.1 Variable: edad 29](#_Toc164932076)

[3.5.2 Variable: costoconsulta 30](#_Toc164932077)

[3.5.3 Variable: costoprocedimiento 31](#_Toc164932078)

[3.5.4 Variable: netoapagarconsulta 32](#_Toc164932079)

[3.5.5 Variable: numerodiaestancia 33](#_Toc164932080)

[3.6 Cobertura geográfica de los datos en la fuente RIPS 34](#_Toc164932081)

[Tabla 1. Descripción inicial de la fuente RIPS 7](#_Toc164932110)

[Tabla 2. Descripción inicial de la fuente RIPS 8](#_Toc164932111)

[Tabla 3. Verificación de diferentes nombres de columnas 9](#_Toc164932112)

[Tabla 4. Muestra de 5 registros duplicados exactamente 10](#_Toc164932113)

[Tabla 5. Cálculo de registros duplicados por año 10](#_Toc164932114)

[Tabla 5. Descripción estadística de las variables​ 12](#_Toc164932115)

[Ilustración 1, Mapa de # registros / población DANE 2023 34](#_Toc164932392)

Imagen 1. Carpetas por año de la fuente RIPS \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_7

[Imagen 2. Error en 24 registros por conteos y tiempos de exportación en la fuente 10](#_Toc164932409)

[Imagen 3. Matriz de correlación de variables numéricas 28](#_Toc164932410)

# Introducción

El presente documento ofrece una visión detallada de los principales hallazgos derivados del proceso de exploración inicial de los datos del sistema RIPS (Registro Individual de Prestación de Servicios de Salud). 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 RIPS, comprende información recopilada durante el periodo comprendido entre 2009 y 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 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.

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 RIPS.

# Metodología

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

## 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, diagramas de bigotes o violín y gráficos de barras, para representar gráficamente la distribución de los datos y las relaciones entre variables.

# Fuente principal de los datos

## Origen y procedencia de los datos

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

|  |  |
| --- | --- |
| Nombre de la fuente en el proyecto: | RIPS |
| Número de archivos o tablas: | 14 carpetas con la información por año |
| Nombre del archivo original de la fuente entregado: | RIPS.zip |
| Ruta del archivo original y archivos de trabajo: | //caoba-access.virtual. .dfs.core.windows.net/rips/rips\_parquet\_{año} |

Tabla 1. Descripción inicial de la fuente RIPS

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 2009 hasta el año 2022.

Tabla

Descripción generada automáticamente  
Imagen 1. Carpeta por años de la fuente RIPS

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

## Descripción general de RIPS y ajustes iniciales

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

|  |  |  |  |
| --- | --- | --- | --- |
| Base | #  de Registros | # de Columnas | ID Únicos |
| RIPS\_2009 | 154.880.279 | 20 | 17.324.543 |
| RIPS\_2010 | 165.813.202 | 20 | 18.639.638 |
| RIPS\_2011 | 224.196.932 | 20 | 20.813.584 |
| RIPS\_2012 | 269.160.205 | 20 | 22.423.812 |
| RIPS\_2013 | 232.756.328 | 20 | 22.277.750 |
| RIPS\_2014 | 317.583.118 | 20 | 26.308.028 |
| RIPS\_2015 | 288.537.345 | 20 | 22.174.604 |
| RIPS\_2016 | 207.392.563 | 20 | 21.052.580 |
| RIPS\_2017 | 280.398.851 | 20 | 24.593.773 |
| RIPS\_2018 | 378.305.881 | 20 | 25.694.513 |
| RIPS\_2019 | 461.480.251 | 21 | 34.657.087 |
| RIPS\_2020 | 409.877.166 | 21 | 29.797.625 |
| RIPS\_2021 | 406.833.792 | 21 | 32.510.848 |
| RIPS\_2022 | 465.644.990 | 21 | 30.441.467 |

Tabla 2. Descripción inicial de la fuente RIPS

El total de registros es de 4’262.860.903 registros en la fuente y 20/21 variables.

### Ajustes requeridos en nombres de las fuentes

La información recibida abarca un periodo que comprende desde 2009 hasta 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 3. Verificación de diferentes nombres de columnas

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

### Ajustes requeridos en registros duplicados

Es importante validar la fuente de datos de registros exactamente iguales o duplicados de todas las columnas. Estos registros en caso de encontrarse deben ser eliminados y dejar solamente un registro.

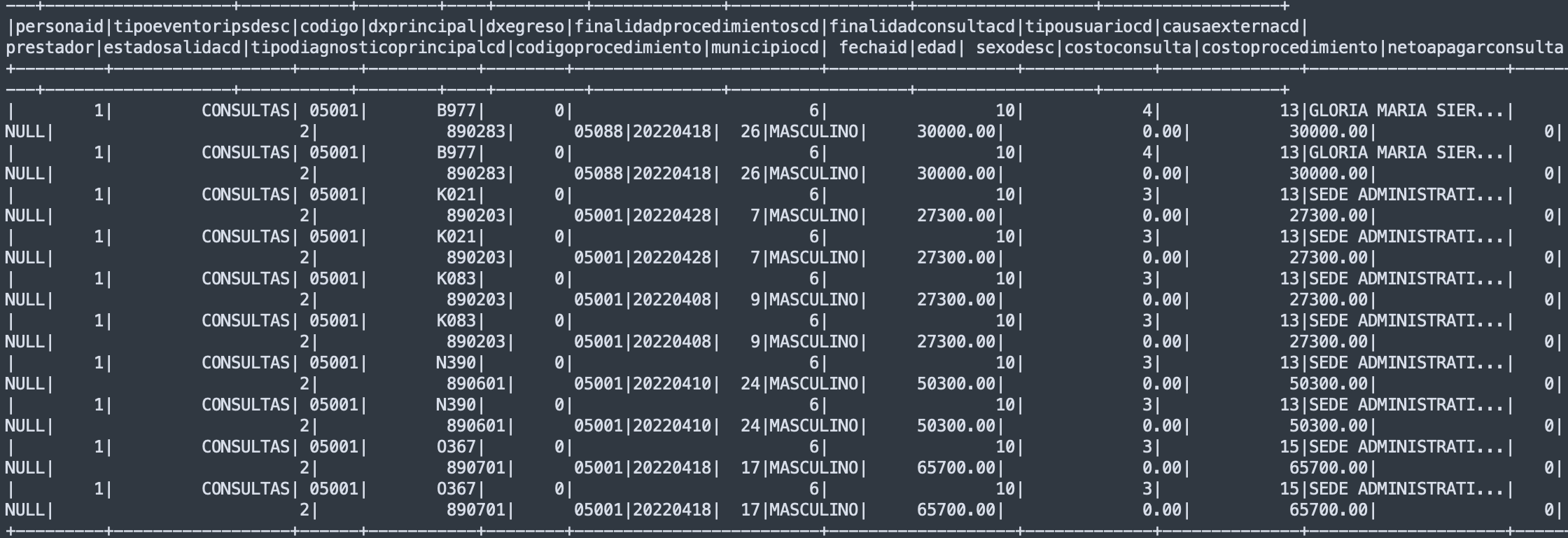


Tabla 4. Muestra de 5 registros duplicados exactamente

Procesando los datos por cada año y validando el número de duplicados se obtiene un total de 323.129 registros, de acuerdo a los siguientes datos calculados.

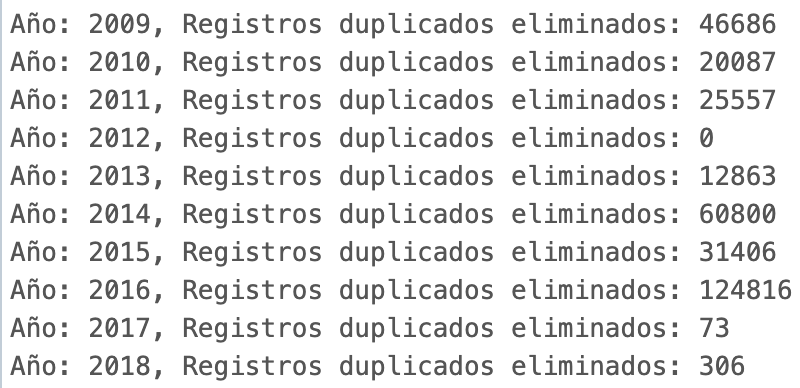
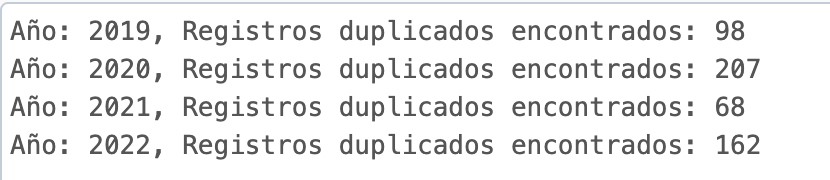
  


Tabla 5. Cálculo de registros duplicados por año

Para efectos del análisis haremos un ajuste en para eliminar los duplicados en la lectura de cada año, y dejaremos un total de registros de 4´262.537.774 para iniciar el análisis descriptivo.

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

Se identificó que la fuente RIPS tiene 24 registros que deben filtrarse/eliminarse antes del análisis descriptivo. Estos registros son el resultado en el proceso de exportación en la fuente de datos donde se dejó el total de registros en todos los años y en algunos años el tiempo empleado del proceso de exportación.

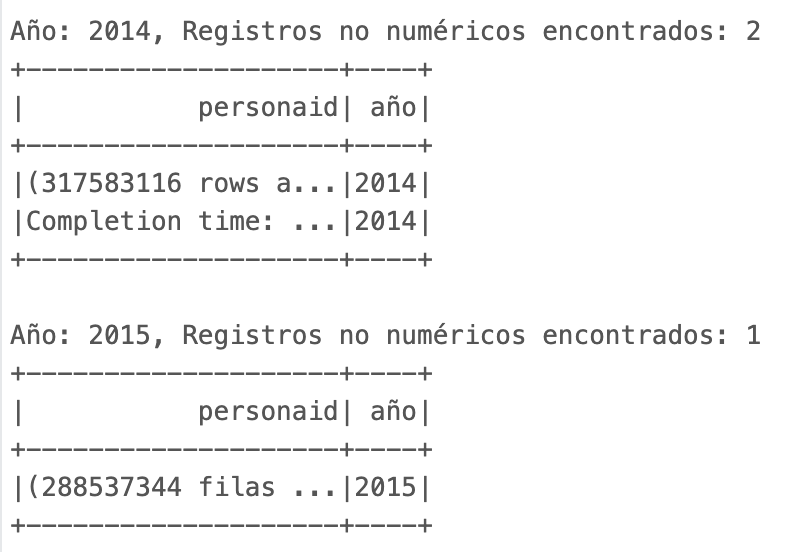


Imagen 2. Error en 24 registros por conteos y tiempos de exportación en la fuente

Luego de este filtro el total de registros de la fuente se disminuyó en 24 registros quedando un total de 4´262.537.750

## Estadísticas descriptivas de las variables

En esta fase, se realizará el análisis exploratorio de los datos de la fuente RIPS. 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.

### 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 RIPS 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 RIPS** | 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 5. Descripción estadística de las variables​

### 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 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 | |
| **Campo 1** | | personaid | |
| Alias | | ID de la persona o paciente | |
| Descripción | | Corresponde al identificador del paciente anonimizado | |
| Tipo | | Categórica nominal | |
| Dominio | | [ENTERO POSITIVO]  Existen 73´357.470 identificadores únicos | |
| Estadísticas Básicas | |  | |
| Observaciones | | * Existen 271.116 registros (1) en la variable * Existe una personaid = 7’747.066, con 359.863 registros | |
| **Campo 2** | | tipoeventoripsdesc | |
| Alias | | Tipo de evento RIPS | |
| Descripción | | Sección (Consulta, Procedimiento, Hospitalización) | |
| Tipo | | Categórico Nominal Alfanumérico | |
| Dominio | | [H,U,HOSPITALIZACIONES,URGENCIAS,C,P,CONSULTAS, PROCEDIMIENTOS DE…]  Existen 8 códigos únicos | |
| Estadísticas Básicas | |  | |
| **Campo 3** | | codigo | |
| Alias | | Código asociado a la consulta | |
| Descripción | | Entidades administradoras de la atención | |
| Tipo | | Categórico Nominal Alfanumérico | |
| Dominio | | [ALFANUMERICOS]  Existen 277 códigos únicos | |
| Estadísticas Básicas | |  | |
| Observaciones | | * Se generó una tabla pivot en Excel de conteo por año y cada código único de esta variable | |
| **Campo 4** | | dxprincipal | |
| Alias | | Código del diagnóstico principal | |
| Descripción | | Código de diagnóstico según la Clasificación Internacional de Enfermedades (CIE-10). Según la sección, tiene una descripción específica.  -Consulta: La afección principal se define como la diagnosticada al final del proceso de atención de la salud o de la consulta, como la causante primaria de la solicitud de atención del paciente.  -Procedimiento: Sólo se diligencia cuando se trata de procedimientos quirúrgicos.  -Hospitalización: Corresponde al diagnóstico que dio origen a la hospitalización: en urgencias, consulta externa o remisión. | |
| Tipo | | Categórico Nominal Alfanumérico | |
| Dominio | | [ALFANUMERICO de 4 posiciones]  Existen 12.645 códigos únicos  Código fuente 240, número registro 3970, 4082 (aunque específica para CIE10) | |
| Estadísticas Básicas | |  | |
|  | | * Se tienen 1´068.224.828 registros con (1) en esta variable * Se generó una tabla pivot en Excel de conteo por año y cada código único de esta variable | |
| **Campo 5** | | dxegreso | |
| Alias | | Código del diagnóstico de egreso | |
| Descripción | | Código de diagnóstico según la Clasificación Internacional de Enfermedades (CIE-10). Debe ser un diagnóstico confirmado en la mayoría de los casos. | |
| Tipo | | Categórico Nominal Alfanumérico | |
| Dominio | | [ALFANUMERICO de 4 posiciones]  Existen 11.400 códigos únicos | |
| Estadísticas Básicas | |  | |
| Observaciones | | * Se tienen 4´233.134.750 registros con (0) en esta variable * Se generó una tabla pivot en Excel de conteo por año y cada código único de esta variable | |
| **Campo 6** | | finalidadprocedimientoscd | |
| Alias | | Código finalidad del procedimiento | |
| Descripción | | Código asociado al objetivo del procedimiento | |
| Tipo | | Numérico Nominal | |
| Dominio | | [0,1,2,3,4,5,6]  Existen 7 códigos únicos  1 = Diagnóstico 2 = Terapéutico 3 = Protección específica 4 = Detección temprana de enfermedad general 5 = Detección temprana de enfermedad laboral | |
| Estadísticas Básicas | |  | |
| Observaciones | | * Se tienen 268 registros con valor (0) | |
| **Campo 7** | | finalidadconsultacd | |
| Alias | | Código finalidad consulta | |
| Descripción | | Código asociado al objetivo de la consulta | |
| Tipo | | Numérico Nominal | |
| Dominio | | [01,02,03,04,05,06,07,08,09,10]  Existen 10 códigos únicos  01 = Atención del parto (puerperio) 02 = Atención del recién nacido 03 = Atención en planificación familiar 04 = Detección de alteraciones de crecimiento y desarrollo del menor de diez años 05 = Detección de alteración del desarrollo joven 06 = Detección de alteraciones del embarazo 07 = Detección de alteraciones del adulto 08 = Detección de alteraciones de agudeza visual 09 = Detección de enfermedad profesional 10 = No aplica | |
| Estadísticas Básicas | |  | |
| Observaciones | | * Se tienen 4´010.148.192 registros en el código (10) en esta variable | |
| **Campo 8** | | tipousuariocd | |
| Alias | | Código asociado al tipo de usuario | |
| Descripción | | Código asociado al tipo de usuario | |
| Tipo | | Numérico Nominal | |
| Dominio | | [1,2,3,4,5,6,7,8]  Existen 8 códigos únicos  1 =Contributivo, 2 =Subsidiado, 3 =Vinculado, 4 =Particular, 5 =Otro, 6 =Víctima con afiliación al Régimen Contributivo, 7 =Víctima con afiliación al Régimen subsidiado, 8 =Víctima no asegurado (Vinculado) | |
| Estadísticas Básicas | |  | |
| Observaciones | | * Se tienen 10’037.671 registros NULL en esta variable | |
| **Campo 9** | | causaexternacd | |
| Alias | | Código asociado a causa externa | |
| Descripción | | Código asociado a causa externa | |
| Tipo | | Numérico Nominal | |
| Dominio | | [01,02,03,04,05,06,07,08,09,10,11,12,13,14,15]  Existen 15 códigos únicos  01 = Accidente de trabajo 02 = Accidente de tránsito 03 = Accidente rábico 04 = Accidente ofídico 05 = Otro tipo de accidente 06 = Evento catastrófico 07 = Lesión por agresión 08 = Lesión auto infligida 09 = Sospecha de maltrato físico 10 = Sospecha de abuso sexual 11 = Sospecha de violencia sexual 12 = Sospecha de maltrato emocional 13 = Enfermedad general 14 = Enfermedad laboral 15 = Otra) | |
| Estadísticas Básicas | |  | |
| Observaciones | | * Se tienen 2´583.938.215 registros NULL en esta variable | |
| **Campo 10** | | prestador | |
| Alias | | Lugar donde se presta el servicio | |
| Descripción | | Lugar donde se presta el servicio | |
| Tipo | | Numérico Nominal | |
| Dominio | | [Enteros positivos]  Existen 70.006 códigos únicos | |
| Estadísticas Básicas | |  | |
| Observaciones | | * Se tienen 46´040.007 registros con valor (0) en esta variable * Se generó una tabla pivot en Excel de conteo por año y cada código único de esta variable | |
| **Campo 11** | | estadosalidacd | |
| Alias | | Estado de salida | |
| Descripción | | Estado del paciente a su salida | |
| Tipo | | Numérico Nominal | |
| Dominio | | [1,2]  Existen 2 códigos únicos  1 = Vivo (a), 2 = Muerto (a) | |
| Estadísticas Básicas | |  | |
| Observaciones | | * Existen 4’162.890.667 registros NULL en esta variable | |
| **Campo 12** | | tipodiagnosticoprincipalcd | |
| Alias | | Código del diagnóstico principal | |
| Descripción | | Tipo del diagnóstico principal | |
| Tipo | | Numérico Nominal | |
| Dominio | | [1,2,3]  Existen 3 códigos únicos  1 = Impresión diagnóstica 2 = Confirmado nuevo 3=Confirmado repetido) | |
| Estadísticas Básicas | |  | |
| Observaciones | | * Se tienen 2´691.283.163 registros NULL en esta variable | |
| **Campo 13** | | codigoprocedimiento | |
| Alias | | Código asociado al procedimiento | |
| Descripción | | Código del procedimiento según corresponda en la Clasificación Única de Procedimientos en Salud (CUPS) de Colombia. | |
| Tipo | | Numérico Nominal | |
| Dominio | | [Enteros positivos]  Existen 12.061 códigos únicos | |
| Estadísticas Básicas | |  | |
| Observaciones | | * Se tienen 100´849.317 registros con valor (1) en esta variable * Se generó una tabla pivot en Excel de conteo por año y cada código único de esta variable | |
| **Campo 14** | | municipiocd | |
| Alias | | Código asociado al municipio | |
| Descripción | | Código asociado al municipio. El lugar donde la persona lleva 6 ó más meses viviendo de manera permanente. Se registra el código correspondiente al municipio en la tabla Municipios. | |
| Tipo | | Categórico | |
| Dominio | | [Enteros positivos - Códigos DANE de municipios de Colombia]  Existen 1.139 códigos únicos | |
| Estadísticas Básicas | |  | |
| Observaciones | | * Existen 158´867.838 registros NULL en esta variable * Bogotá D.C. tiene 751´332.379 registros * Medellín tiene 255´369.565 registros | |
| **Campo 15** | | fechaid | |
| Alias | | Fecha asociada | |
| Descripción | | Fecha en se llevó a cabo la atención. | |
| Tipo | | Fecha | |
| Dominio | | [AAAAMMDD]  Existen 5.113 registros únicos | |
| Estadísticas Básicas | |  | |
| **Campo 16** | | edad | |
| Alias | | Edad del usuario o paciente | |
| Descripción | | Se refiere a la edad del usuario en valores enteros en el momento que se realiza la atención. | |
| Tipo | | Numérico continua ordinal cuantitativa | |
| Dominio | | [-8 : 365]  Existen 168 valores únicos | |
| Estadísticas Básicas | |  | |
| Observaciones | | * Existen 11 registros con valores de edad negativos * Count (Recuento): Hay 4,262,537.750 registros en la columna edad. * Mean (Media): La edad promedio es de 40.86 años, lo que sugiere que la población es relativamente joven en general. * StdDev (Desviación Estándar): La desviación estándar es de 24.31 años, lo que indica una variabilidad significativa en las edades. * Min (Mínimo): La edad mínima es -8.0, lo que parece ser un valor anómalo o incorrecto, ya que no es posible tener una edad negativa. * Max (Máximo): La edad máxima es 365.0, lo que también parece ser un valor anómalo o incorrecto, ya que supera los límites de la expectativa de vida humana. | |
| **Campo 17** | | sexodesc | |
| Alias | | Sexo del paciente | |
| Descripción | | Sexo del usuario o paciente | |
| Tipo | | Categórico Alfanumérico | |
| Dominio | | [FEMENINO, MASCULINO,F,M,NO DEFINIDO,N,INDEFINIDO,INDETERMINADO]  Existen 8 categorías únicas | |
| Estadísticas Básicas | |  | |
| Observaciones | | * Se tienen 18´.578.562 registros NULL en esta variable. * Se deben reclasificar categorias (F y M) y (FEMENINO y MASCULINO) * Determinar si las categorias N, NO DEFINIDO, INDETERMINADO y INDEFINIDO pueden agruparse en una sola categoría | |
| **Campo 18** | | costoconsulta | |
| Alias | | Costo asociado a la consulta | |
| Descripción | | Costo asociado a la consulta | |
| Tipo | | Numérico | |
| Dominio | | [-8´458.800 : 1.531448 E14]  Existen 275.211 valores únicos | |
| Estadísticas Básicas | |  | |
| Observaciones | | * Count (Recuento): Hay 4,262,537,750 registros en la columna costoconsulta. * Mean (Media): El costo promedio de consulta es de aproximadamente 2,317,703.62. * StdDev (Desviación Estándar): La desviación estándar es muy alta, lo que indica una variabilidad significativa en los costos de consulta. * Min (Mínimo): El costo mínimo es negativo (-8,458,800.0), lo cual puede ser un error o un reembolso. * Max (Máximo): El costo máximo es extremadamente alto (1.531448E14), lo cual puede ser un error o un dato atípico. | |
| **Campo 19** | | costoprocedimiento | |
| Alias | | Costo asociado al procedimiento | |
| Descripción | | Costo asociado al procedimiento | |
| Tipo | | Numérico | |
| Dominio | | [-3´270.000 : 8.009583 E14]  Existen 1´659.365 valores únicos | |
| Estadísticas Básicas | |  | |
| Observaciones | | * Count (Recuento): Hay 4,262,537,750 registros en la columna costoprocedimiento. * Mean (Media): El costo promedio de procedimiento es de aproximadamente 3,770,369.09. * StdDev (Desviación Estándar): La desviación estándar es extremadamente alta, lo que indica una variabilidad significativa en los costos de procedimiento. * Min (Mínimo): El costo mínimo es negativo (-3,270,000.0), lo cual puede ser un error o un reembolso. * Max (Máximo): El costo máximo es extremadamente alto (8.009583E14), lo cual puede ser un error o un dato atípico. | |
| **Campo 20** | | netoapagarconsulta | |
| Alias | | Cantidad neta a pagar por la consulta | |
| Descripción | | Este dato corresponde al valor que el prestador cobrará al pagador por esa consulta. | |
| Tipo | | Numérico | |
| Dominio | | [0 : 2.3462799 E14]  Existen 327.400 valores únicos | |
| Estadísticas Básicas | |  | |
| Observaciones | | * Count (Recuento): Hay 1´743.835.656 registros en la columna netoapagarconsulta. * Esta variable solamente está para los años 2019 a 2022 * Mean (Media): El neto a pagar promedio es de aproximadamente 6,096,855.20. * StdDev (Desviación Estándar): La desviación estándar es muy alta, lo que indica una gran variabilidad en los valores. * Min (Mínimo): El valor mínimo es 0, lo que parece correcto si los valores negativos se consideraron como reembolsos. * Max (Máximo): El valor máximo es extremadamente alto (2.3462799E14), lo que podría ser un dato atípico. | |
| **Campo 21** | | numerodiasestancia | |
| Alias | | Cantidad de días de estancia | |
| Descripción | | Cantidad de días de estancia | |
| Tipo | | Numérico | |
| Dominio | | [0 : 15.480]  Existen 2.085 valores únicos | |
| Estadísticas Básicas | |  | |
| Observaciones | | * Count (Recuento): Hay 4,262,537,750 registros en la columna numerodiasestancia. * Mean (Media): El número promedio de días de estancia es muy bajo (0.0376), lo que indica que la mayoría de los registros tienen muy pocos días de estancia. * StdDev (Desviación Estándar): La desviación estándar es baja, lo que indica poca variabilidad en el número de días de estancia. * Min (Mínimo): El número mínimo de días de estancia es 0, lo que parece correcto. * Max (Máximo): El número máximo de días de estancia es 15,408, lo cual parece muy alto y podría ser un dato atípico. | |

### Análisis preliminares

**Variables numéricas**

* Los datos presentan una variabilidad en costos y días de estancia, lo que se puede esperar en los registros médicos manejados en el tiempo.
* Hay valores atípicos o incorrectos, como por ejemplo la edad negativa, costos muy altos o negativos, los cuales podrían ser investigados para verificar la calidad de los datos y constatar que sean cuentas de alto costo.
* En general, los datos muestran una población relativamente joven (40.8) con costos y días de estancia que pueden variar

**Variables categóricas**

* Hay una concentración de registros en ciertas personas o identificadores, lo que sugiere múltiples atenciones a una misma persona.
* En el tipo de evento, los registros se centran en procedimientos, consultas y urgencias, con hospitalizaciones siendo menos frecuentes.
* Hay una gran variedad de los códigos de salud, con algunos códigos más predominantes, indicando concentración en ciertas actividades de salud.
* En el diagnostico principal hay una variedad de los diagnósticos, con algunos más comunes que otros, reflejando problemas de salud prevalentes en la población.
* Los tipos de usuarios más comunes son 1, 2, y 5, indicando una concentración en ciertos tipos de usuarios.
* Hay una concentración en ciertos prestadores de servicios de salud, indicando posibles relaciones con ciertas instituciones.
* Algunos códigos de procedimiento como 890201 y 890301 son predominantes, indicando procedimientos comunes en la población.
* Los datos están distribuidos a lo largo de diversas fechas, sugiriendo una cobertura temporal amplia, con algunas fechas en particulares con más registros.

## Análisis de correlación

Se realizó una exploración de la relación entre pares de variables numéricas mediante el cálculo de matriz de correlación, como el coeficiente de correlación de Pearson. Con el objetivo de identificar posibles asociaciones lineales o no lineales entre las variables.

### ¿Qué es el Coeficiente de Correlación?

El coeficiente de correlación mide la relación lineal entre dos variables numéricas. El coeficiente de correlación de Pearson, que es comúnmente utilizado y probablemente lo que se calcula en la matriz, varía entre -1 y +1. Aquí está cómo interpretar estos valores:

* +1: Una correlación perfectamente positiva. Esto significa que a medida que una variable aumenta, la otra también aumenta en una relación lineal perfecta.
* -1: Una correlación perfectamente negativa. Significa que a medida que una variable aumenta, la otra disminuye en una relación lineal perfecta.
* 0: No hay correlación lineal. Las dos variables no se influyen entre sí de manera lineal.

Cómo resultado se obtiene la matriz de correlación mostrada en la siguiente imagen:

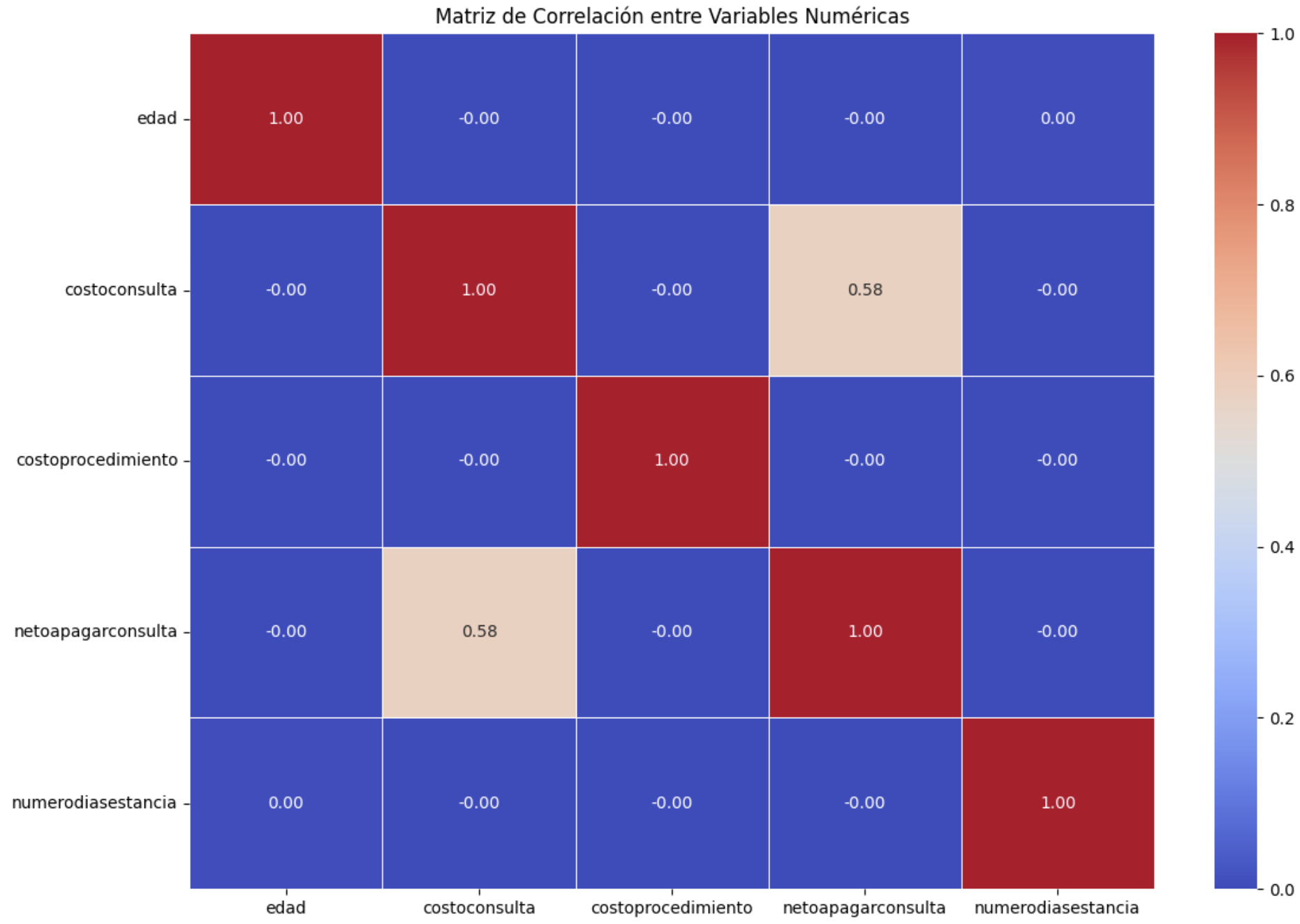


Imagen 3. Matriz de correlación de variables numéricas

* Diagonal: La diagonal principal (de arriba a la izquierda a abajo a la derecha) siempre muestra 1.00, porque es la correlación de cada variable consigo misma.
* Costo de Consulta y Neto a Pagar por Consulta: Un valor de 0.58 indica una correlación moderada positiva entre estas dos variables. Esto sugiere que, en general, a mayor costo de la consulta, mayor es el neto a pagar por la consulta.
* Otros Valores de 0.00 o cercanos a cero: Estos valores indican que no hay una correlación lineal detectable entre las variables comparadas. Por ejemplo, la edad no tiene correlación lineal con el costo de consulta, costo de procedimiento, neto a pagar, o número de días de estancia, según los datos.

## Visualización de datos

Se generaron gráficos y visualizaciones, como histogramas, y diagramas de bigotes y gráficos de barras, para representar gráficamente la distribución de los datos. Debido al tamaño al tamaño de la fuente existen varias técnicas para hacer estas visualizaciones en Datos BigData.

Para este caso con las variables numéricas se tomó una muestra del 10% de los datos.

### Variable: edad

|  |  |  |
| --- | --- | --- |
| Escala lineal | * Rango de Edades: Las edades se extienden desde 0 hasta más de 200, lo cual es poco común y sugiere posibles errores de entrada de datos o valores atípicos, ya que las edades superiores a 120 son extremadamente raras. * Frecuencias: La mayoría de las edades se concentran en el rango inferior del espectro, con la frecuencia más alta en las primeras barras del histograma, que parecen incluir las edades desde el nacimiento hasta alrededor de los 50-60 años. * Forma de la Distribución: La distribución de las edades parece sesgada hacia la derecha, indicando que hay más registros de poblaciones más jóvenes y disminuye a medida que aumenta la edad. * Escala de Frecuencia: La escala vertical está en notación científica (por ejemplo, 1e7), lo que indica el número de registros para cada rango de edad. Por ejemplo, una barra que llegue a 2.5e7 indica 25 millones de registros para ese rango de edad. * Valores Anómalos: Las barras que se extienden más allá de la edad típica de vida humana sugieren la presencia de valores anómalos o errores en los datos. | * La mediana está cerca de la parte inferior de la caja, lo que podría indicar que la distribución de las edades es asimétrica con una cola más larga hacia los valores más altos. * Hay varios valores atípicos que están muy por encima del tercer cuartil y los bigotes. Esto podría deberse a errores de entrada de datos, a una población con una cantidad significativa de personas de edad avanzada, o a otros factores específicos del conjunto de datos. |
| Escala logarít mica | * Concentración de Datos: La mayoría de los datos se concentran alrededor del rango de 1.5 en la escala logarítmica. Esto indica que hay una edad específica (o un rango de edades) que es más común en tu conjunto de datos, después de aplicar la transformación logarítmica. * Cauda a la Derecha: Hay una cola larga hacia valores más altos de log\_edad, lo que indica que hay menos ocurrencias de valores más altos de edad, pero todavía son significativas. | * Mediana: La línea verde en el centro de la caja representa la mediana de la distribución de la log\_edad. Está situada alrededor de 1.0 en la escala logarítmica, indicando que el valor de la mediana de la edad transformada cae alrededor de este punto. * Cuartiles: El borde inferior de la caja representa el primer cuartil (Q1), y el borde superior representa el tercer cuartil (Q3). El espacio entre estos dos representa el rango intercuartílico (IQR), que contiene la mitad central de los datos. La caja no parece ser simétrica alrededor de la mediana, lo que sugiere cierta asimetría en la distribución de la log\_edad. * Bigotes (Whiskers): Los bigotes representan el rango del resto de los datos, exceptuando valores atípicos, y en este gráfico, no se extienden demasiado por encima del tercer cuartil o por debajo del primer cuartil, indicando que los valores extremos de log\_edad no están demasiado lejos del IQR. * Valores Atípicos (Outliers): Los círculos debajo del primer cuartil indican valores atípicos en el extremo inferior de la distribución. Estos pueden representar edades particularmente bajas después de la transformación logarítmica. |

### Variable: costoconsulta

|  |  |  |
| --- | --- | --- |
| Escala logarítmica | * Pico Pronunciado: Existe un pico pronunciado en los valores bajos del logaritmo del costo de consulta, lo que sugiere que la mayoría de los costos de consulta están en el rango bajo después de la transformación logarítmica. * Distribución Asimétrica: La distribución parece ser asimétrica con una cola larga que se extiende hacia valores más altos de logaritmo del costo de consulta. Esto indica que hay menos registros con un costo de consulta alto, pero aún son significativos. * Valores Cercanos a Cero: Hay una frecuencia relativamente alta de valores cercanos a cero en el logaritmo del costo de consulta, lo que puede indicar una presencia considerable de consultas con costos muy bajos o cercanos a cero antes de la transformación logarítmica. | * Mediana: La línea azul claro dentro del cuadro indica la mediana de los datos, y parece estar por debajo de 2 en la escala logarítmica. Esto sugiere que la mitad de los valores de costo de consulta están por debajo de este punto en la escala logarítmica. * Cuartiles: El cuadro (la caja del boxplot) muestra el rango intercuartílico (IQR), que contiene el 50% central de los datos. La posición del cuadro sugiere que la mayoría de los costos de consulta se concentran en un rango relativamente bajo en la escala logarítmica. * Bigotes (Whiskers): Los extremos de las líneas verticales (los bigotes) representan aproximadamente los valores mínimo y máximo dentro de un rango aceptable (normalmente 1.5 veces el IQR desde la parte superior e inferior de la caja). Dado que el bigote inferior está muy cerca del cuadro, indica que hay una menor variabilidad en los valores bajos de costo de consulta. * Valores atípicos (Outliers): Los círculos representan valores atípicos, es decir, valores que caen fuera del rango de los bigotes. Hay un número significativo de valores atípicos que son considerablemente más altos que el rango típico del costo de consulta. Estos puntos sugieren que hay casos de costos de consulta que son inusualmente altos en comparación con la mayoría. * Escala Logarítmica: Dado que el gráfico está en una escala logarítmica, las diferencias en los valores altos representan cambios exponenciales en los costos de consulta. Incluso pequeñas diferencias en la escala logarítmica pueden significar grandes diferencias en los valores originales. |

### Variable: costoprocedimiento

|  |  |  |
| --- | --- | --- |
| Escala logarítmica | * Concentración de Datos: Hay una clara concentración de datos alrededor del valor 2 en la escala logarítmica, lo que indica que la mayoría de los costos de procedimiento se agrupan en torno a un rango específico después de aplicar la transformación logarítmica. * Rango de Valores: Los costos de procedimiento cubren un rango desde valores muy bajos hasta aproximadamente el valor 8 en la escala logarítmica, lo que sugiere que existe una variabilidad significativa en los costos de procedimiento. * Cauda Larga a la Derecha: El histograma tiene una "cola" larga que se extiende hacia la derecha, indicando que hay menos ocurrencias de costos de procedimiento más altos. Esta es una característica común en los datos financieros donde muchas observaciones están agrupadas en torno a valores más bajos con algunos valores extremadamente altos (esto se conoce como distribución sesgada a la derecha). | * Mediana: La línea horizontal azul claro dentro de la caja indica la mediana de los datos, que parece estar alrededor de 4 en la escala logarítmica. * Cuartiles: El borde inferior de la caja representa el primer cuartil (Q1), y el borde superior representa el tercer cuartil (Q3), lo que implica que el 50% central de los datos se encuentra entre estos dos puntos. El rango intercuartílico (IQR) parece ser estrecho, indicando que la mitad central de los datos se agrupa dentro de un rango limitado de valores en la escala logarítmica. * Bigotes (Whiskers): Los extremos de las líneas verticales (bigotes) que se extienden desde la caja indican la variabilidad fuera de los cuartiles superior e inferior. No se extienden mucho más allá de la caja, lo que sugiere que la mayoría de los datos están contenidos dentro de este rango. * Valores Atípicos (Outliers): Los círculos representan valores que caen fuera del rango definido por los bigotes. Algunos puntos están significativamente por encima del tercer cuartil, lo que indica la presencia de costos de procedimiento muy altos en comparación con el resto. |

### Variable: netoapagarconsulta

|  |  |  |
| --- | --- | --- |
| Escala logarítmica | * Pico Pronunciado: Existe un pico muy pronunciado alrededor del valor 4 en la escala logarítmica, lo que indica una concentración alta de registros en este valor específico o en un rango estrecho alrededor de él. * Cauda a la Derecha: Aunque hay una caída drástica en la frecuencia después del pico, hay una cola larga hacia la derecha del histograma. Esto sugiere que hay algunos valores que son significativamente más altos que la mayoría de las observaciones, indicando la presencia de consultas con montos netos a pagar relativamente altos. * Asimetría: La distribución es asimétrica, con un sesgo hacia la derecha. En una escala logarítmica, esto puede sugerir que el monto neto a pagar por la consulta tiene un rango que se extiende hacia valores altos, pero con menos frecuencia que los valores más bajos. | * Mediana: La línea azul claro dentro de la caja indica la mediana de los datos, y parece estar cerca de 2.5 en la escala logarítmica. Esto significa que la mitad de los valores del monto neto a pagar por las consultas está por debajo de este valor y la otra mitad por encima, después de aplicar la transformación logarítmica. * Cuartiles: La caja representa el rango intercuartílico (IQR), delimitado por el primer cuartil (Q1) en la parte inferior y el tercer cuartil (Q3) en la parte superior. La concentración de datos dentro del IQR es relativamente estrecha, lo que indica que una gran parte de los montos netos a pagar por las consultas en su forma logarítmica cae dentro de un rango pequeño de valores. * Bigotes (Whiskers): Los bigotes indican la variabilidad fuera de los cuartiles superior e inferior y se extienden hasta el valor máximo y mínimo dentro de un rango aceptable, que suele ser de 1.5 veces el IQR desde la parte superior e inferior de la caja. La presencia de valores atípicos sugiere que se supera este rango. * Valores Atípicos (Outliers): Los círculos fuera de los bigotes indican valores atípicos, que son valores que caen fuera del rango definido por 1.5 veces el IQR. Estos valores atípicos son notablemente más altos que el Q3, indicando la presencia de montos netos a pagar por las consultas que son inusualmente altos, incluso después de la transformación logarítmica. |

### Variable: numerodiaestancia

|  |  |  |
| --- | --- | --- |
| Escala logarítmica | * Mayor Frecuencia en Valores Bajos: La mayoría de los registros tienen valores bajos en la escala logarítmica, lo que sugiere que hay una alta frecuencia de estancias cortas. * Disminución Gradual: A medida que el valor del logaritmo aumenta, la frecuencia de registros disminuye de manera más o menos gradual, lo que indica que las estancias más largas son menos comunes. * Cauda hacia la Derecha: Existe una cauda que se extiende hacia valores más altos en la escala logarítmica, aunque la frecuencia de estos valores es baja en comparación con los valores más bajos. Esto es típico de los datos de estancia donde la mayoría de las estancias son cortas, pero hay unas pocas mucho más largas. * No hay Barras en Valores Más Altos: No parece haber frecuencias significativas más allá de un cierto punto en la escala logarítmica, indicando un límite en la duración de las estancias que se presentan en el conjunto de datos. | * Mediana: La línea azul en el centro del cuadro indica la mediana de los datos en la escala logarítmica, que parece estar alrededor de 0.5. * Rango Intercuartílico (IQR): La caja define el IQR y muestra dónde se encuentra la mitad central de los datos. La posición de la caja sugiere que la mayoría de los días de estancia, en su forma logarítmica, se agrupan alrededor de valores bajos. * Bigotes: Los extremos de las líneas verticales (bigotes) indican el rango de los datos fuera del IQR, excluyendo valores atípicos. El bigote superior se extiende más lejos de la caja que el inferior, lo que sugiere que hay más variabilidad en el extremo superior de los días de estancia, incluso después de la transformación logarítmica. * Valores Atípicos: Los círculos por encima del bigote superior representan valores atípicos. En la escala logarítmica, estos valores son significativamente más altos que el Q3 y podrían representar estancias anormalmente largas. |

Al interpretar este tipo de visualizaciones, especialmente cuando se utilizan transformaciones logarítmicas, es importante recordar que estamos trabajando con una representación distorsionada de los datos. Esto puede ser útil para identificar patrones y tendencias en datos con un amplio rango de valores, pero se debe tener precaución al volver a convertir estos valores a la escala original, ya que las diferencias relativas pueden ser exponencialmente mayores. Además, los valores atípicos en una escala logarítmica pueden indicar variaciones extremas en la escala original, y pueden ser dignos de un análisis más detallado para entender sus causas.

## Cobertura geográfica de los datos en la fuente RIPS

La fuente de datos ubica geográficamente los registros en la fuente RIPS. En la siguiente imagen se presenta un mapa que permite determinar de forma normalizada (de acuerdo con población 2023 según el DANE) la distribución de registros de la fuente RIPS en los departamentos de Colombia.

Es importante tener cuidado en esta lectura, dado que las atenciones de usuarios de acuerdo con el tipo de atención pueden generar diferente número de registros por cada paciente.

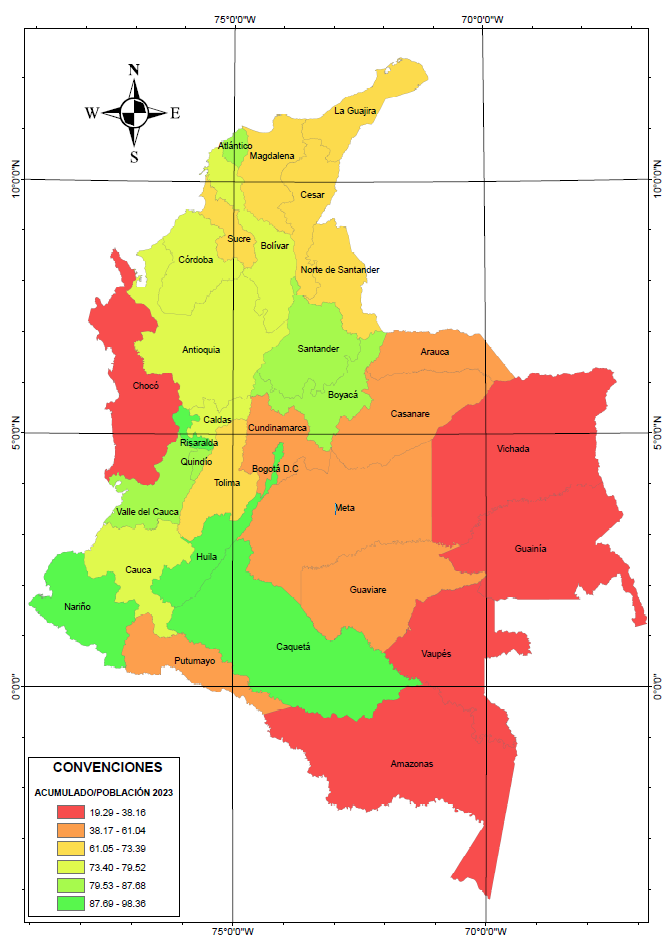


Ilustración 1, Mapa de # registros / población DANE 2023

Fuente: elaborado por los autores