Imagen que contiene Forma

Descripción generada automáticamente

**PROYECTO AGORA**

**ENTENDIMIENTO DE LOS DATOS - SEGCOVID**

**DOCUMENTO INICIAL**

**2024**

**Control de cambios**

|  |  |  |  |
| --- | --- | --- | --- |
| **VERSIÓN** | **FECHA** | **MODIFICACIÓN** | **RESPONSABLE** |
| 1.3 | 24 de abril de 2024 | Modificación del documento | Luis Carlos Chicaíza |
| 1.5 | 07 de mayo de 2025 | Ajustes del documento con base a los comentarios recibidos | Luis Carlos Chicaíza |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Contenido

[1. Introducción 5](#_Toc165995106)

[2. Metodología 6](#_Toc165995107)

[2.1 Exploración de los datos 6](#_Toc165995108)

[3. Fuente principal de los datos 7](#_Toc165995109)

[3.1 Origen y procedencia de los datos 7](#_Toc165995110)

[3.2 Descripción general de SEGCOVID 8](#_Toc165995111)

[3.2.1 Ajustes requeridos en nombres de las fuentes 9](#_Toc165995112)

[3.2.2 Ajustes requeridos en registros de resumen en la exportación 9](#_Toc165995113)

[3.3 Estadísticas descriptivas de las variables 11](#_Toc165995114)

[3.3.1 Esquema para la exploración y análisis descriptivo de los datos 12](#_Toc165995115)

[3.3.2. Listado detallado de la fuente 13](#_Toc165995116)

[3.3.3 Análisis preliminares 38](#_Toc165995117)

[3.4 Análisis de correlación 39](#_Toc165995118)

Ilustraciones

[Imagen 1. Carpetas por años de la fuente SEGCOVID 7](#_Toc164970968)

[Imagen 2. Valores no numéricos 9](#_Toc164970969)

[Imagen 3. Datos nulos 10](#_Toc164970970)

## Introducción

El presente documento ofrece una visión detallada de los principales hallazgos derivados del proceso de exploración inicial de los datos del sistema SEGCOVID. 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 SEGCOVID, comprende información recopilada durante el periodo comprendido entre 2020 a 2023. 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, 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.

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 SEGCOVID.

## Metodología

Con el propósito de realizar un análisis exhaustivo de los datos provenientes de la fuente SEGCOVID 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: | SEGCOVID |
| Número de archivos o tablas1: | 10 carpetas con la información por año |
| Nombre del archivo original de la fuente entregado2: | Segcovid.zip |
| Ruta del archivo original y archivos de trabajo3: | //caoba-access.virtual. .dfs.core.windows.net/segcovid/segcovid\_parquet\_{año} |

Tabla 1. Descripción inicial de la fuente SEGCOVID

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 2020 hasta el año 2023.

Tabla

Descripción generada automáticamente

Imagen 1. Carpetas por años de la fuente SEGCOVID

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 SEGCOVID

A continuación, se detalla los registros por año de los registros proporcionados:

|  |  |  |  |
| --- | --- | --- | --- |
| Base | #  de Registros | # de Columnas | ID Únicos |
| SEGCOVID\_2020 | 1.923.699 | 40 | 1.834.066 |
| SEGCOVID\_2021 | 13.882.989 | 40 | 3.141.047 |
| SEGCOVID\_2022 | 2.566.855 | 40 | 749.120 |
| SEGCOVID\_2023 | 58.190 | 40 | 33.563 |

Tabla 2. Descripción inicial de la fuente SEGCOVID

El total de registros es de 18’431.733 registros en el seguimiento de los casos de COVID-19 conocidos como SEGCOVID.

### 3.2.1 Ajustes requeridos en nombres de las fuentes

La información recibida abarca un periodo que comprende desde 2020 hasta 2023. 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.

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

Al analizar los valores no numéricos de los datos originales de SEGCOVID se identificó que todas las variables tienen el mismo número de registros y son categóricas.

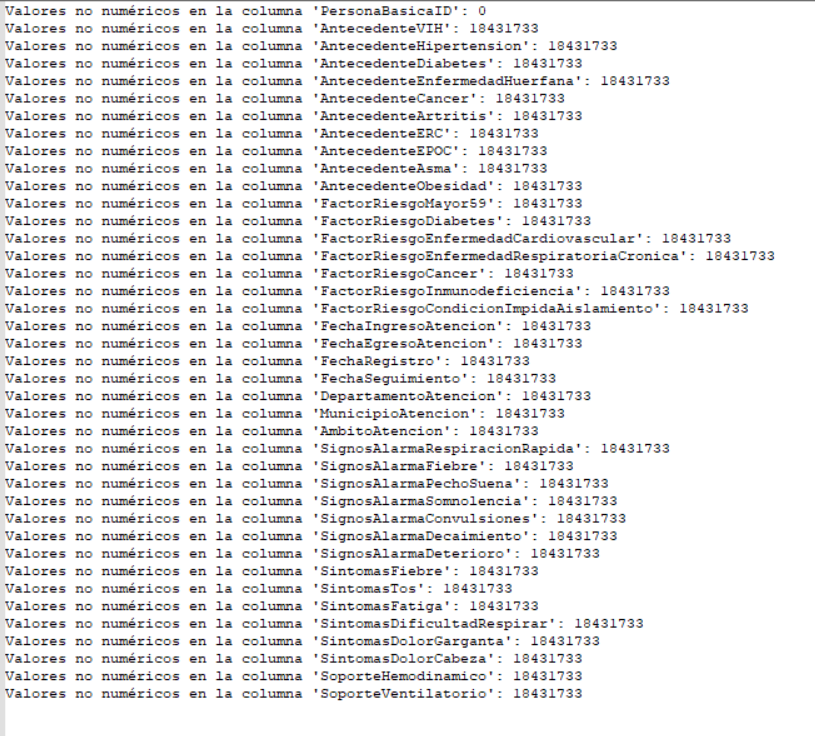


Imagen 2. Valores no numéricos

Al comprobar estos registros se validó si había datos en nulo, no se evidencia ningún registro con esta característica.

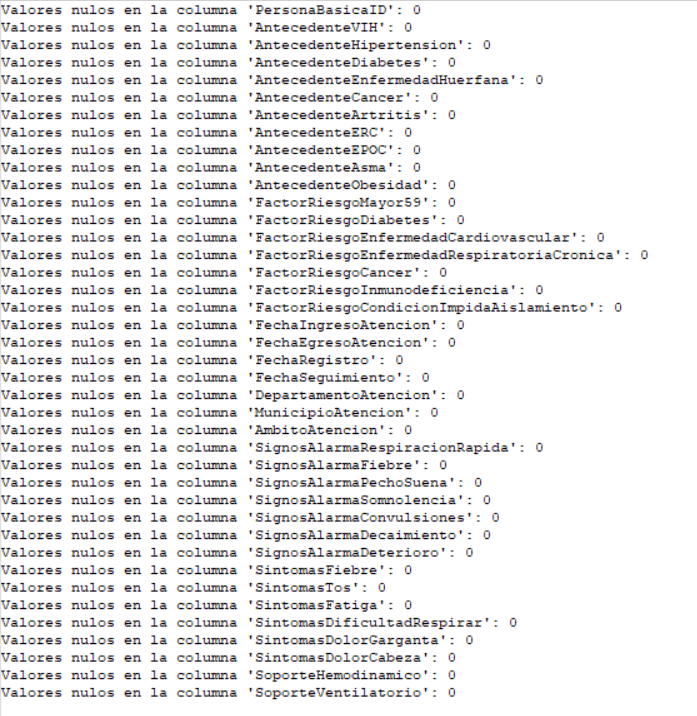


Imagen 3. Datos nulos

Por lo tanto, no se modificó la fuente original y sigue con los 18’431.733 registros.

## 3.3 Estadísticas descriptivas de las variables

En esta fase se realizará el análisis exploratorio de los datos de la fuente SEGCOVID. 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 SEGCOVID 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 SEGCOVID** | 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 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 |
| **Campo 1** | PersonaBasicaID |
| Alias | ID de la persona |
| Descripción | Corresponde al identificador del paciente anonimizado. |
| Tipo | Categórico Nominal |
| Dominio | [ENTERO POSITIVO]  Existen 5’633.748 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Existe una PersonaID = 96’647.755, con 589 registros    Los siguientes más comunes son:   * 138647957 con 579 registros * 46316220 con 570 registros * 1 con 530 registros * 37551922 con 410 registros |
| **Campo 2** | AntecedenteVIH |
| Alias | Pacientes con antecedentes de VIH |
| Descripción | Describe si los pacientes tienen antecedentes de VIH |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no tienen antecedentes de VIH, con 18,362,377 casos. * Solo hay 69,356 registros con antecedentes de VIH. |
| **Campo 3** | **AntecedenteHipertension** |
| Alias | Pacientes con antecedentes de Hipertensión |
| Descripción | Describe si los pacientes tienen antecedentes de Hipertensión. |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no tienen antecedentes de hipertensión, con 16,122,338 casos. * Hay 2,309,395 registros con antecedentes de hipertensión. |
| **Campo 4** | AntecedenteDiabetes |
| Alias | Pacientes con antecedentes de Diabetes |
| Descripción | Describe si los pacientes tienen antecedentes de Diabetes |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no tienen antecedentes de diabetes, con 17,649,079 casos. * Hay 782,654 registros con antecedentes de diabetes. |
| **Campo 5** | AntecedenteEnfermedadHuerfana |
| Alias | Pacientes con antecedentes de enfermedad huérfana |
| Descripción | Describe si los pacientes tienen antecedentes de enfermedad huérfana |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no tienen antecedentes de enfermedad huérfana, con 18,372,541 casos. * Hay 59,192 registros con antecedentes de enfermedad huérfana. |
| **Campo 6** | AntecedenteCancer |
| Alias | Pacientes con antecedentes de Cáncer |
| Descripción | Describe si los pacientes tienen antecedentes Cáncer |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no tienen antecedentes de cáncer, con 18,231,865 casos. * Hay 199,868 registros con antecedentes de cáncer. |
| **Campo 7** | AntecedenteArtritis |
| Alias | Pacientes con antecedentes de Artritis |
| Descripción | Describe si los pacientes tienen antecedentes de Artritis |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no tienen antecedentes de artritis, con 18,373,009 casos. * Hay 58,724 registros con antecedentes de artritis. |
| **Campo 8** | AntecedenteERC |
| Alias | Pacientes con antecedentes de ERC |
| Descripción | Describe si los pacientes tienen antecedentes de ERC |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no tienen antecedentes de enfermedad renal crónica, con 17,969,459 casos. * Hay 462,274 registros con antecedentes de enfermedad renal crónica. |
| **Campo 9** | AntecedenteEPOC |
| Alias | Pacientes con antecedentes de EPOC |
| Descripción | Describe si los pacientes tienen antecedentes de EPOC |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no tienen antecedentes de EPOC, con 17,828,598 casos. * Hay 603,135 registros con antecedentes de EPOC. |
| **Campo 10** | AntecedenteAsma |
| Alias | Pacientes con antecedentes de asma |
| Descripción | Describe si los pacientes tienen antecedentes de |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no tienen antecedentes de asma, con 17,881,745 casos. * Hay 549,988 registros con antecedentes de asma. |
| **Campo 11** | AntecedenteObesidad |
| Alias | Pacientes con antecedentes de obesidad |
| Descripción | Describe si los pacientes tienen antecedentes de obesidad |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no tienen antecedentes de obesidad, con 16,013,225 casos. * Hay 2,418,508 registros con antecedentes de obesidad. |
| **Campo 12** | FactorRiesgoMayor59 |
| Alias | Factor de riesgo para mayor de 59 años |
| Descripción | Describe si los pacientes mayores de 59 años tienen riesgo |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no presentan factor de riesgo mayor a 59 años, con 16,224,535 casos. * Hay 2,207,198 registros con factor de riesgo mayor a 59 años. |
| **Campo 13** | FactorRiesgoDiabetes |
| Alias | Factor de riesgo para diabetes |
| Descripción | Describe si los pacientes tienen riesgo de diabetes |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no presentan factor de riesgo asociado a diabetes, con 18,084,955 casos. * Hay 346,778 registros con factor de riesgo asociado a diabetes. |
| **Campo 14** | FactorRiesgoEnfermedadCardiovascular |
| Alias | Factor de riesgo de enfermedad cardiovascular |
| Descripción | Describe si los pacientes tienen riesgo de enfermedad cardiovascular |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no presentan factor de riesgo asociado a enfermedad cardiovascular, con 17,512,702 casos. * Hay 919,031 registros con factor de riesgo asociado a enfermedad cardiovascular. |
| **Campo 15** | FactorRiesgoEnfermedadRespiratoriaCronica |
| Alias | Factor de riesgo para enfermedad respiratoria crónica |
| Descripción | Describe si los pacientes tienen riesgo de enfermedad respiratoria crónica |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no presentan factor de riesgo asociado a enfermedad respiratoria crónica, con 18,259,751 casos. * Hay 171,982 registros con factor de riesgo asociado a enfermedad respiratoria crónica. |
| **Campo 16** | FactorRiesgoCancer |
| Alias | Factor de riesgo para cáncer |
| Descripción | Describe si los pacientes tienen riesgo de cáncer |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no presentan factor de riesgo asociado a cáncer, con 18,307,785 casos. * Hay 123,948 registros con factor de riesgo asociado a cáncer. |
| **Campo 17** | FactorRiesgoInmunodeficiencia |
| Alias | Factor de riesgo para inmunodeficiencia |
| Descripción | Describe si los pacientes tienen riesgo de inmunodeficiencia |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no presentan factor de riesgo asociado a inmunodeficiencia, con 18,379,099 casos. * Hay 52,634 registros con factor de riesgo asociado a inmunodeficiencia. |
| **Campo 18** | FactorRiesgoCondicionImpidaAislamiento |
| Alias | Factor de riesgo para condición impida aislamiento |
| Descripción | Describe si los pacientes tienen riesgo de condición impida aislamiento |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no presentan factor de riesgo asociado a una condición que impida el aislamiento, con 18,394,062 casos. * Hay 37,671 registros con factor de riesgo asociado a una condición que impida el aislamiento. |
| **Campo 19** | FechaIngresoAtencion |
| Alias | Fecha ingreso atención |
| Descripción | Describe la fecha de ingreso del paciente para la atención |
| Tipo | Categórico Nominal/ Fecha |
| Dominio | [AAAA-MM-DD]  Existen 1161 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La fecha “1900-01-01”, puede ser un valor predeterminado para indicar una fecha desconocida o no especificada. |
| **Campo 20** | FechaEgresoAtencion |
| Alias | Fecha egreso atención |
| Descripción | Describe la fecha de egreso del paciente de la atención |
| Tipo | Categórico Nominal/ Fecha |
| Dominio | [AAAA-MM-DD]  Existen 1138 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La fecha “1900-01-01”, puede ser un valor predeterminado para indicar una fecha desconocida o no especificada. |
| **Campo 21** | FechaRegistro |
| Alias | Fecha registro |
| Descripción | Describe la fecha de registro del paciente |
| Tipo | Categórico Nominal/ Fecha |
| Dominio | [AAAA-MM-DD]  Existen 1019 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros tienen fechas de registro en 2021 y 2022. * La fecha de registro más común es "2021-06-28" con 1,412,905 registros. |
| **Campo 22** | FechaSeguimiento |
| Alias | Fecha seguimiento |
| Descripción | Describe la fecha de seguimiento del paciente |
| Tipo | Categórico Nominal/ Fecha |
| Dominio | [AAAA-MM-DD]  Existen 1792 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La fecha “1900-01-01”, puede ser un valor predeterminado para indicar una fecha desconocida o no especificada. |
| **Campo 23** | DepartamentoAtencion |
| Alias | Departamento de atención |
| Descripción | Describe el departamento de atención |
| Tipo | Categórico Nominal |
| Dominio | [Código – Nombre Departamento]  Existen 34 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros provienen de Bogotá D.C. y otras áreas urbanas, como Antioquia, Valle del Cauca y Atlantico. * La mayoría de los registros están relacionados con atención en el departamento de Bogotá D.C., con 5,910,399 casos. |
| **Campo 24** | MunicipioAtencion |
| Alias | Municipio de atención |
| Descripción | Describe el municipio de atención |
| Tipo | Categórico Nominal |
| Dominio | [Código – Nombre Municipio]  Existen 1123 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros están relacionados con atención en Bogotá D.C., con 5,910,399 casos. * Hay otros municipios de atención presentes en los datos. |
| **Campo 25** | AmbitoAtencion |
| Alias | Ámbito de atención |
| Descripción | Describe el ámbito de atención |
| Tipo | Categórico Nominal |
| Dominio | [No Reportado, Hospitalización, Cuidado Intensivo, Cuidado Intermedio, Urgencias, Domiciliaria, Consulta Externa, Otro]  Existen 8 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no tienen un ámbito de atención reportado, con 17,483,602 casos. * Hay otros ámbitos de atención presentes en los datos, como hospitalización, cuidado intensivo, cuidado intermedio, urgencias, etc. |
| **Campo 26** | SignosAlarmaRespiracionRapida |
| Alias | Signos de alarma de respiración rápida |
| Descripción | Describe si hay signos de alarma por respiración rápida |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no tienen signos de alarma de respiración rápida, con 18,147,993 casos. * Hay 283,740 registros con signos de alarma de respiración rápida. |
| **Campo 27** | SignosAlarmaFiebre |
| Alias | Signos de alarma fiebre |
| Descripción | Describe si hay signos de alarma por fiebre |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no tienen signos de alarma de fiebre, con 18,241,111 casos. * Hay 190,622 registros con signos de alarma de fiebre. |
| **Campo 28** | SignosAlarmaPechoSuena |
| Alias | Signos de alarma de pecho suena |
| Descripción | Describe si hay signos de alarma por pecho |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no tienen signos de alarma de pecho que suena, con 18,235,786 casos. * Hay 195,947 registros con signos de alarma de pecho que suena. |
| **Campo 29** | SignosAlarmaSomnolencia |
| Alias | Signos de alarma somnolencia |
| Descripción | Describe si hay signos de alarma por somnolencia |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no tienen signos de alarma de somnolencia, con 18,397,347 casos. * Hay 34,386 registros con signos de alarma de somnolencia. |
| **Campo 30** | SignosAlarmaConvulsiones |
| Alias | Signos de alarma convulsiones |
| Descripción | Describe si hay signos de alarma por convulsiones |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no tienen signos de alarma de convulsiones, con 18,428,661 casos. * Hay 3,072 registros con signos de alarma de convulsiones. |
| **Campo 31** | SignosAlarmaDecaimiento |
| Alias | Signos de alarma decaimiento |
| Descripción | Describe si hay signos de alarma por decaimiento |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no tienen signos de alarma de decaimiento, con 18,429,835 casos. * Hay 1,898 registros con signos de alarma de decaimiento. |
| **Campo 32** | SignosAlarmaDeterioro |
| Alias | Signos de alarma deterioro |
| Descripción | Describe si hay signos de alarma por deterioro |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no tienen signos de alarma de deterioro, con 18,278,297 casos. * Hay 153,436 registros con signos de alarma de deterioro. |
| **Campo 33** | SintomasFiebre |
| Alias | Síntomas fiebre |
| Descripción | Describe si hay síntomas de fiebre |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no tienen síntomas de fiebre, con 16,913,621 casos. * Hay 1,518,112 registros con síntomas de fiebre. |
| **Campo 34** | SintomasTos |
| Alias | Síntomas tos |
| Descripción | Describe si hay síntomas de tos |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no tienen síntomas de tos, con 15,126,146 casos. * Hay 3,305,587 registros con síntomas de tos. |
| **Campo 35** | SintomasFatiga |
| Alias | Síntomas fatiga |
| Descripción | Describe si hay síntomas de fatiga |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no tienen síntomas de fatiga, con 15,767,606 casos. * Hay 2,664,127 registros con síntomas de fatiga. |
| **Campo 36** | SintomasDificultadRespirar |
| Alias | Síntomas dificultad respirar |
| Descripción | Describe si hay síntomas por dificultad de respirar |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no tienen síntomas de dificultad para respirar, con 17,892,647 casos. * Hay 539,086 registros con síntomas de dificultad para respirar. |
| **Campo 37** | SintomasDolorGarganta |
| Alias | Síntomas dolor de garganta |
| Descripción | Describe si hay síntomas por dolor de garganta |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no tienen síntomas de dolor de garganta, con 16,384,278 casos. * Hay 2,047,455 registros con síntomas de dolor de garganta. |
| **Campo 38** | SintomasDolorCabeza |
| Alias | Síntomas dolor de cabeza |
| Descripción | Describe si hay síntomas por dolor de cabeza |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros no tienen síntomas de dolor de cabeza, con 17,554,428 casos. * Hay 877,305 registros con síntomas de dolor de cabeza. |
| **Campo 39** | SoporteHemodinamico |
| Alias | Soporte Hemodinámico |
| Descripción | Describe si el paciente requiere un soporte hemodinámico |
| Tipo | Categórico Nominal |
| Dominio | [Si, No]  Existen 2 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los pacientes no requirieron soporte ventilatorio. |
| **Campo 40** | SoporteVentilatorio |
| Alias | Soporte Ventilatorio |
| Descripción | Describe si el paciente requiere un tipo de soporte ventilatorio. |
| Tipo | Categórico Nominal |
| Dominio | [No Reportado, Cánula nasal, IOT, Máscara, CPAP, Ventury,Ninguno]  Existen 7 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los pacientes no reportan solicitar un soporte ventilatorio. * El soporte mas solicitado es la Cánula nasal con 264.113 casos. |

Tabla 4: Descripción datos fuente

### 3.3.3 Análisis preliminares

**Variables categóricas**

* Hay una gran cantidad de valores únicos para ‘PersonaBasicaID’ (5,633,748).
* Las columnas relacionadas con antecedentes muestran solo dos valores únicos, lo que significa que son variables categóricas con solo dos posibles estados (Si o No). En todos los casos, el estado "No" es significativamente más común que el estado "Sí".
* 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.
* Las columnas ‘DepartamentoAtencion’ y ‘MunicipioAtencion’ tienen una variedad de valores únicos, lo que indica diversidad en las ubicaciones de atención. La mayoría de los registros provienen de Bogotá D.C. y otras áreas urbanas, como Medellín, Cali y Barranquilla.
* La columna ‘AmbitoAtencion’ muestra que la mayoría de los casos fueron atendidos en un ámbito no reportado o en hospitalización.
* Varias columnas están asociadas con signos de alarma y síntomas, estas también muestran solo dos valores únicos, indicando la presencia o ausencia de los síntomas. La mayoría de los registros muestran ausencia de síntomas o signos de alarma.
* Los datos muestran una tendencia en que la mayoría de los pacientes no requirieron soporte hemodinámico o ventilatorio.

**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 evidenciaron variables numéricas en la fuente de datos SEGCOVID.
* Para una mayor profundidad de los datos se puede analizar los resultados de la exploración de los datos realizada en databricks.

## 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 todas las variables fueron catalogadas como categóricas.