**Prueba Ingeniero de Datos Nequi**

**Jesús Emilio Valderrama de la Espriella**

**Paso 1: Alcance del proyecto y captura de datos**

Se desea analizar la relación existente entre los fallecidos por COVID-19 en cada uno de los departamentos de Colombia contra los afiliados del régimen subsidiado a cada una de las EPS departamentales o nacionales, con el objetivo de tomar acciones para garantizar la oportuna atención de los ciudadanos.

Se utilizarán las siguientes fuentes de datos

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Nombre** | **Descripción** | **Fuente** | **Cantidad de registros** | **¿Frecuencia de actualización** | **¿Cómo se accede a los datos?** |
| Casos positivos COVID-19 Colombia | Detalle de cada uno de los casos positivos en Colombia incluyendo ubicación y estado de recuperación. | [Ministerio de Salud y Protección Social](https://www.datos.gov.co/Salud-y-Protecci-n-Social/Casos-positivos-de-COVID-19-en-Colombia/gt2j-8ykr/data) | 3.724.705 | Diaria | SODA API desde datos abiertos Colombia |
| Base\_eps\_Colombia | Cantidad de registros por EPS régimen subsidiado por departamento y grupo poblacional | [Ministerio de Salud y Protección Social](https://www.datos.gov.co/Salud-y-Protecci-n-Social/Poblaci-n-Base-de-Datos-nica-de-Afiliados-BDUA-del/d7a5-cnra/data) | 702.820 | Mensual | Descarga manual datos abiertos y almacenamiento en Google Cloud Storage. Se adjunta key para acceder a los datos |

Los datos se desean preparar realizar una tabla de análisis e informe mensual al ministerio Salud y Protección Social.

**Descripción de campos base Casos positivos COVID-19 Colombia**

|  |  |  |
| --- | --- | --- |
| **Nombre de Columna** | **Descripción** | **Tipo** |
| **fecha reporte web** | Fecha de publicación en sitio web | Texto simple |
| **ID de caso** | Identificación del caso | Número |
| **Fecha de notificación** | Fecha de notificación a SIVIGILA | Texto simple |
| **Código DIVIPOLA departamento** | Código Departamento | Número |
| **Nombre departamento** | Nombre Departamento | Texto simple |
| **Código DIVIPOLA municipio** | Código Municipio | Número |
| **Nombre municipio** | Nombre Municipio | Texto simple |
| **Edad** | Edad | Número |
| **Unidad de medida de edad** | Unidad de medida de la edad: 1-Años 2-Meses 3-Días | Número |
| **Sexo** | Género | Texto simple |
| **Tipo de contagio** | Relacionado Importado En estudio Comunitario | Texto simple |
| **Ubicación del caso** | \* Corresponde a muertes no relacionadas con COVID-19, aún si eran casos activos \*\*Hay pacientes recuperados para COVID-19, que pueden permanecer en hospitalización por otras comorbilidades | Texto simple |
| **Estado** | \* Corresponde a muertes no relacionadas con COVID-19, aún si eran casos activos \*\*Hay pacientes recuperados para COVID-19, que pueden permanecer en hospitalización por otras comorbilidades | Texto simple |
| **Código ISO del país** | Código de país de importación | Número |
| **Nombre del país** | Nombre del país de importación | Texto simple |
| **Recuperado** | Recuperado Fallecido N/A (Vacío). N/A se refiere a los fallecidos no COVID. Puede haber casos recuperados con ubicación Hospital u Hospital UCI, ya que permanecen en hospitalización por causas diferentes. Los casos con información en blanco en esta columna corresponden a los casos activos | Texto simple |
| **Fecha de inicio de síntomas** | Fecha de inicio de síntomas | Texto simple |
| **Fecha de muerte** | En caso de ser fallecido | Texto simple |
| **Fecha de diagnóstico** | Fecha de confirmación por laboratorio | Texto simple |
| **Fecha de recuperación** | Fecha de recuperación | Texto simple |
| **Tipo de recuperación** | Se refiere a la variable de tipo de recuperación que tiene dos opciones: PCR y tiempo. PCR indica que la persona se encuentra recuperada por segunda muestra, en donde dio negativo para el virus; mientras que tiempo significa que son personas que cumplieron 30 días posteriores al inicio de síntomas o toma de muestras que no tienen síntomas, que no tengan más de 70 años ni que estén hospitalizados. | Texto simple |
| **Pertenencia étnica** | 1-Indígena 2-ROM 3-Raizal 4-Palenquero 5-Negro 6-Otro. Esta variable se actualizará cada semana. ADVERTENCIA DE RESPONSABILIDAD: La variable etnia depende totalmente de tres cosas: - El correcto diligenciamiento de la variable Etnia por los profesionales de salud que notifican en más de 10.000 instituciones de salud en todos los municipios y departamentos. - Del autorreconocimiento de la persona cuando se le pregunta por esta variable. - Del listado censal que haga y mantenga actualizado cada departamento. No depende del Instituto Nacional de Salud, y por lo tanto, es responsabilidad de las autoridades de cada municipio, departamento y distrito de Colombia; la calidad y consistencia de dicha variable | Número |
| **Nombre del grupo étnico** | ADVERTENCIA DE RESPONSABILIDAD: La variable etnia depende totalmente de tres cosas: - El correcto diligenciamiento de la variable Etnia por los profesionales de salud que notifican en más de 10.000 instituciones de salud en todos los municipios y departamentos. - Del autorreconocimiento de la persona cuando se le pregunta por esta variable. - Del listado censal que haga y mantenga actualizado cada departamento. No depende del Instituto Nacional de Salud, y por lo tanto, es responsabilidad de las autoridades de cada municipio, departamento y distrito de Colombia; la calidad y consistencia de dicha variable | Texto simple |

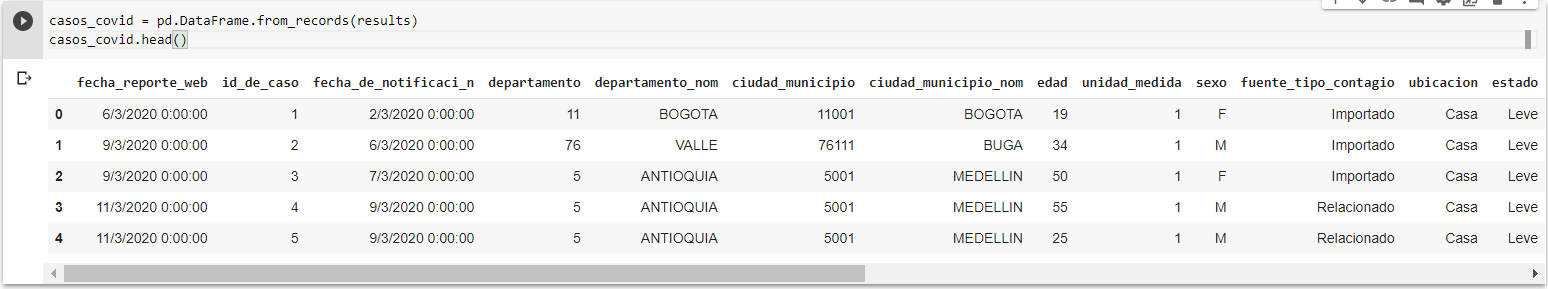
**Descripción de campos Base\_eps\_Colombia.**

|  |  |  |
| --- | --- | --- |
| **Nombre de Columna** | **Descripción** | **Tipo** |
| **Genero** | Valores: Masculino, Femenino | Texto simple |
| **Grupo etario** | Primera Infancia (0-5 años), Infancia (6 - 11 años), Adolescencia (12 - 18 años), Juventud (18 - 26 años), Adultez (27- 59 años), Persona Mayor (60 años o mas) envejecimiento y vejez | Texto simple |
| **Código de la entidad** | Código EPS | Texto simple |
| **Nombre de la entidad** | Nombre EPS | Texto simple |
| **Régimen al que pertenece** | Contributivo, Subsidiado | Texto simple |
| **Tipo de afiliado** | Cotizante, Beneficiario, Adicional, Cabeza de Familia | Texto simple |
| **Estado del afiliado** | Activo, Retirado, Fallecido, Suspendido, Protección Laboral. | Texto simple |
| **Condición del beneficiario** | Urbana, Rural. | Texto simple |
| **Zona de Afiliación** | Zona de Colombia | Texto simple |
| **Departamento** | Departamento | Texto simple |
| **Municipio** | Municipio | Texto simple |
| **Nivel del Sisbén** | I, II, III | Texto simple |
| **Grupo poblacional del afiliado** | Grupo poblacional | Texto simple |
| **cantidad** | Cantidad de afiliados | Texto simple |

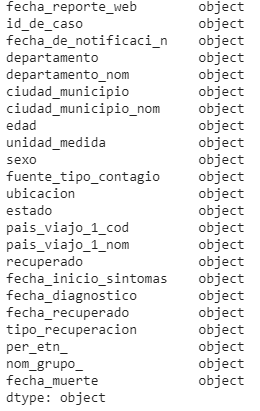
**Paso 2: Explorar y evaluar los datos, el EDA.**

**Exploración de datos Casos positivos COVID-19 Colombia**

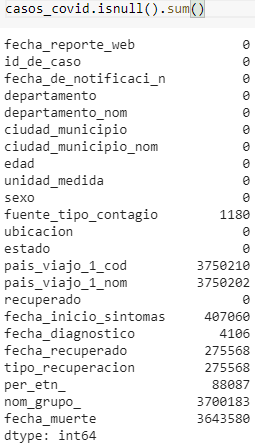
**Registros cargados correctamente desde la API**

****

**Tipos de datos cargados**

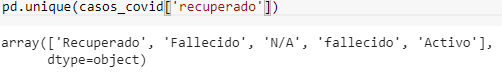
****

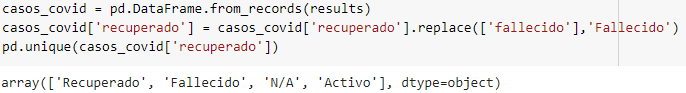
**Identificación de registros nulos**

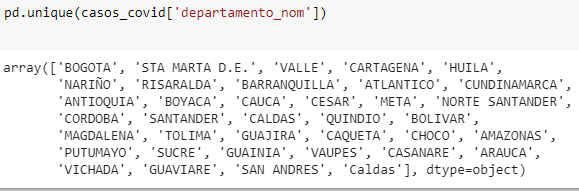
****

El comportamiento de registros nulos es normal, debido a que todas las columnas no se deben diligenciar. Para el análisis que se desea realizar, sólo se tendrá en cuentan los registros fallecidos por departamento y el grupo etario.

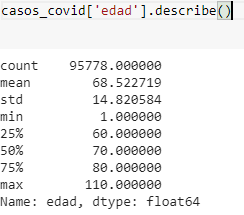
**Estandarización de la columna recuperado**

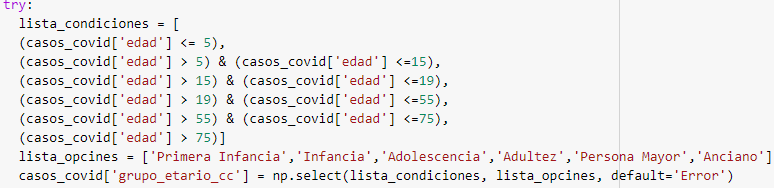
**  
Resultado después de la estandarización para proceder a filtrar**

****

****

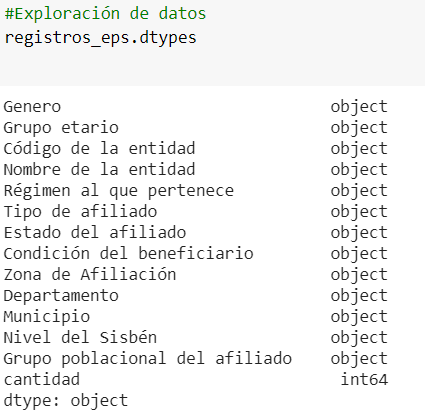
**Se crea una nueva columna para el grupo etario según la edad**

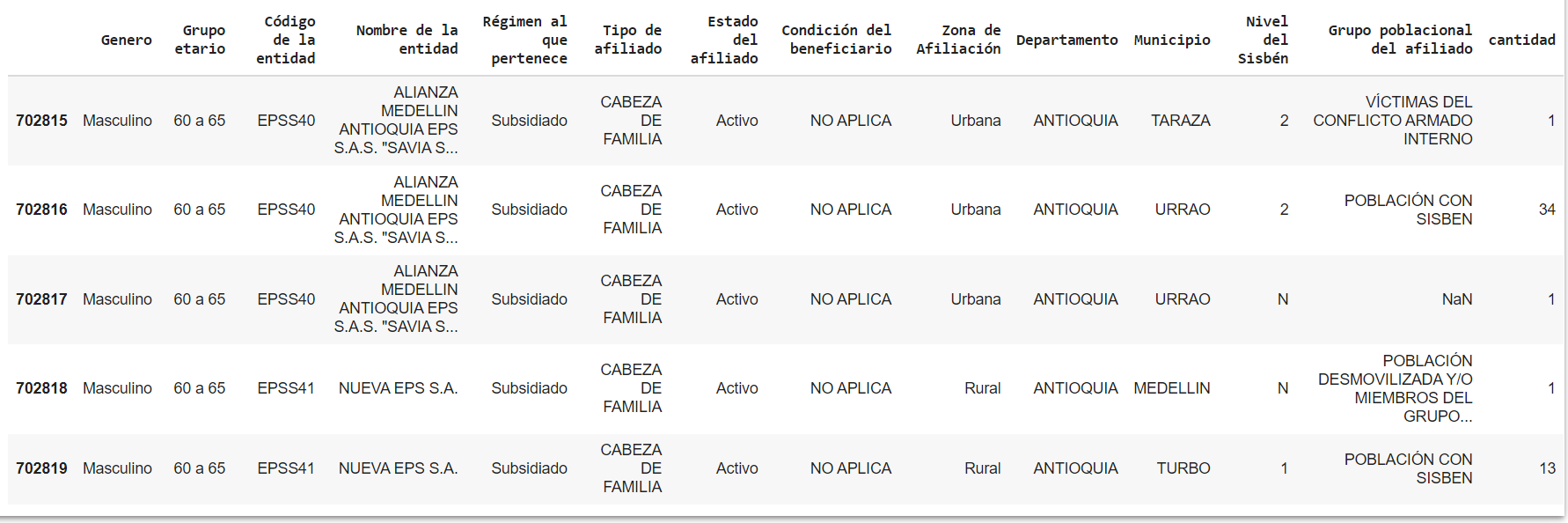
****

****

**Exploración de datos** Base\_eps\_Colombia

Descripción de los tipos de datos



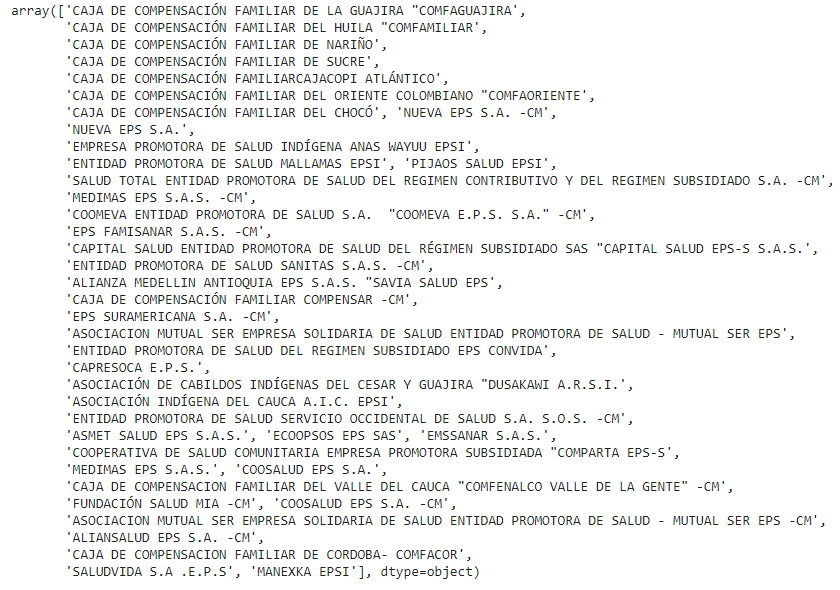


Se realiza cambio de nombre de los campos, con el objetivo de eliminar tildes, espacios y caracteres especiales.

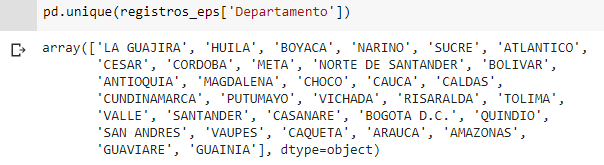
|  |  |
| --- | --- |
| **Nombre de Columna anterior** | **Nuevo nombre** |
| **Genero** | genero |
| **Grupo etario** | grupo\_etario |
| **Código de la entidad** | cod\_entidad |
| **Nombre de la entidad** | nombre\_entidad |
| **Régimen al que pertenece** | regimen |
| **Tipo de afiliado** | tipo\_afiliado |
| **Estado del afiliado** | estado\_afiliado |
| **Condición del beneficiario** | condicion\_benef |
| **Zona de Afiliación** | zona\_afiliacion |
| **Departamento** | departamento |
| **Municipio** | municipio |
| **Nivel del Sisbén** | nivel\_sisben |
| **Grupo poblacional del afiliado** | grupo\_poblacional |
| **cantidad** | cantidad\_afiliados |



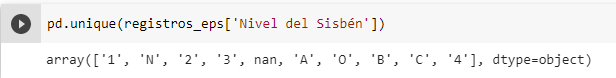
Identificación de registros únicos campo eps

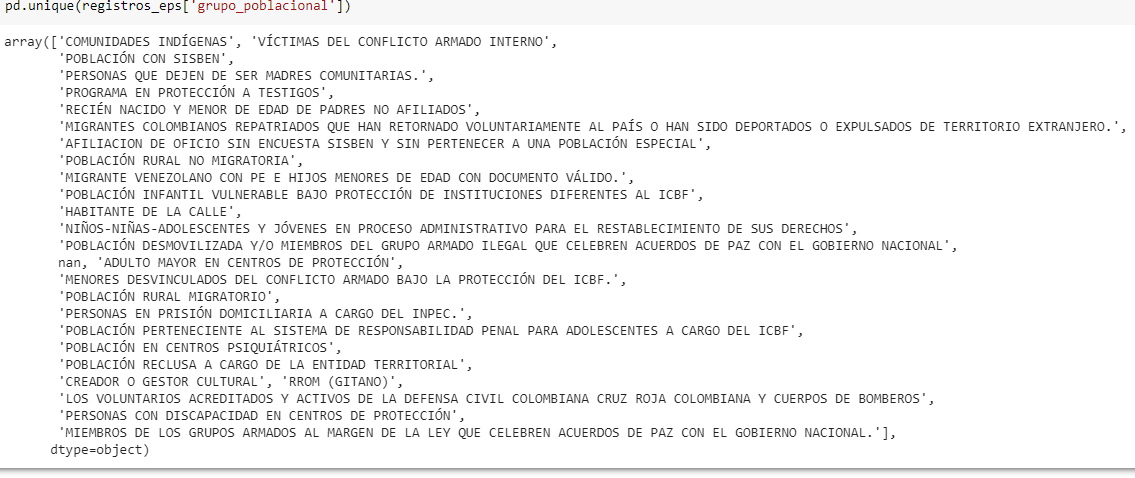


Identificación registros únicos departamentos, esto con el objetivo de validar si es necesario homologar nombres.

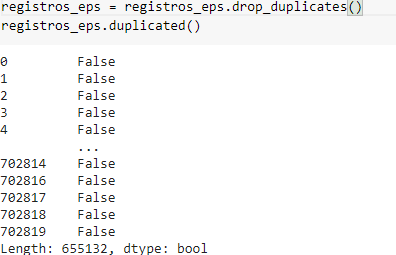






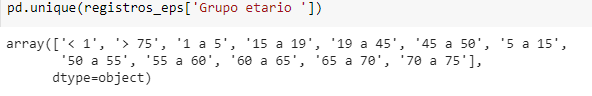


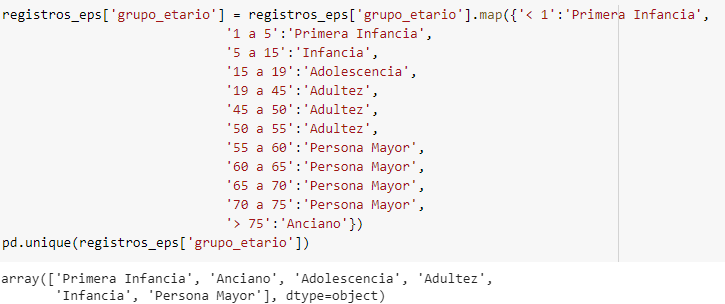
Se eliminan registros duplicados



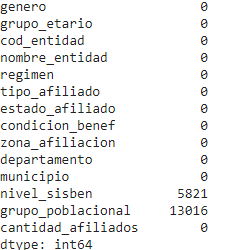
Para la columna “Grupo etario “, se realizará una limpieza de datos y se homologará de acuerdo con lo estipulado por el ministerio de Salud y Protección Social, así:

Primera Infancia (0-5 años), Infancia (6 - 11 años), Adolescencia (12 - 18 años), Juventud (18 - 26 años), Adultez (27- 59 años), Persona Mayor (60 años o mas) envejecimiento y vejez



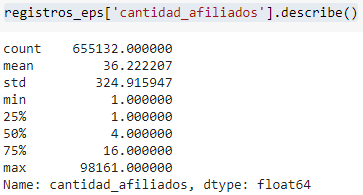


Se identifican las columnas con registros nulos

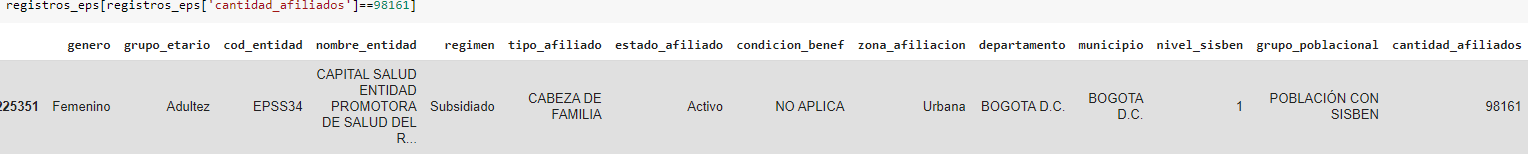


No se realiza ningún tratamiento a los registros nulos, debido a que para el análisis deseado no son necesarios las columnas nivel\_sisben y grupo poblacional.

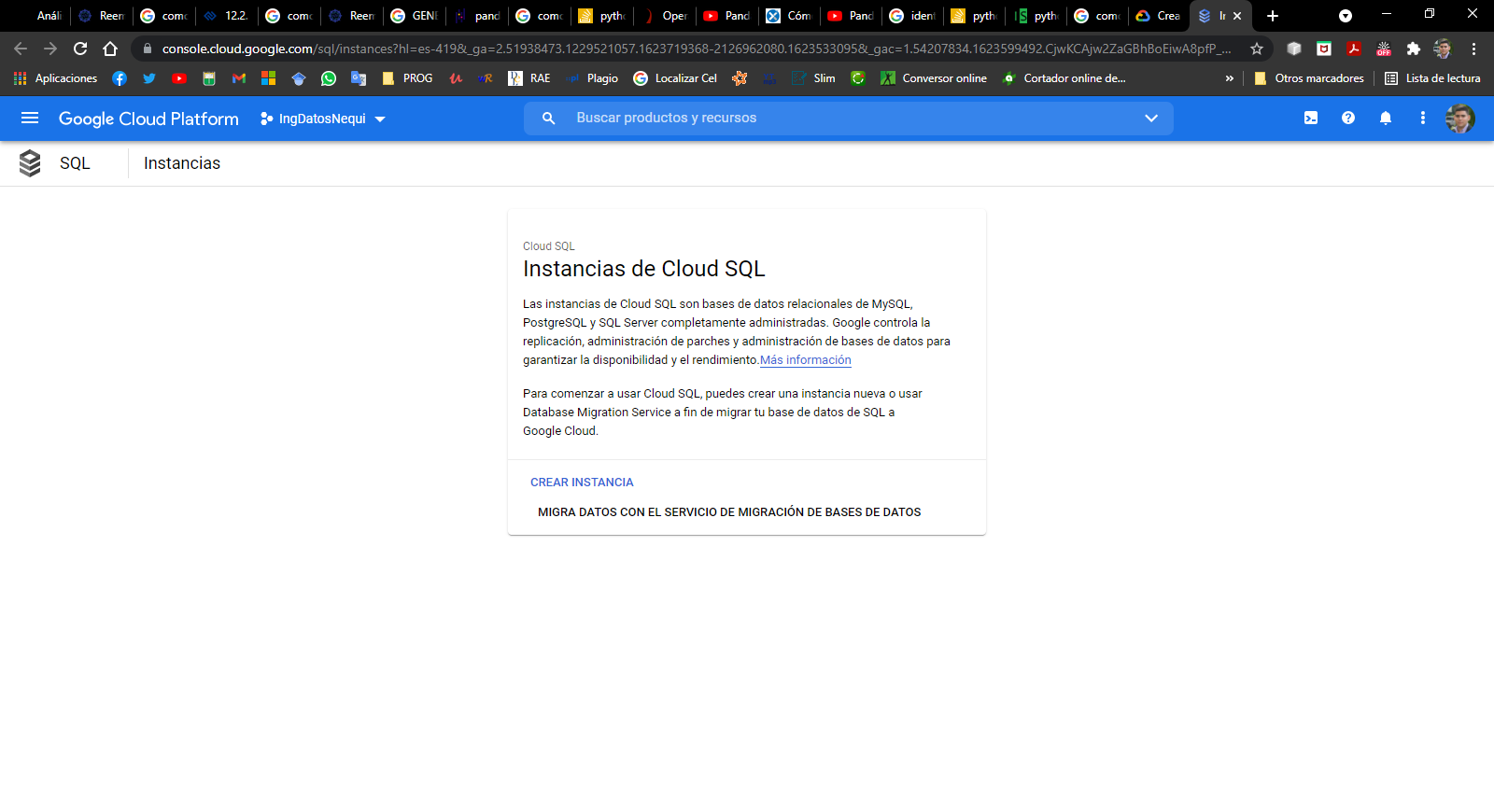
Se realiza análisis estadístico básico para identificar atípicos en el campo cantidad\_afiliados.



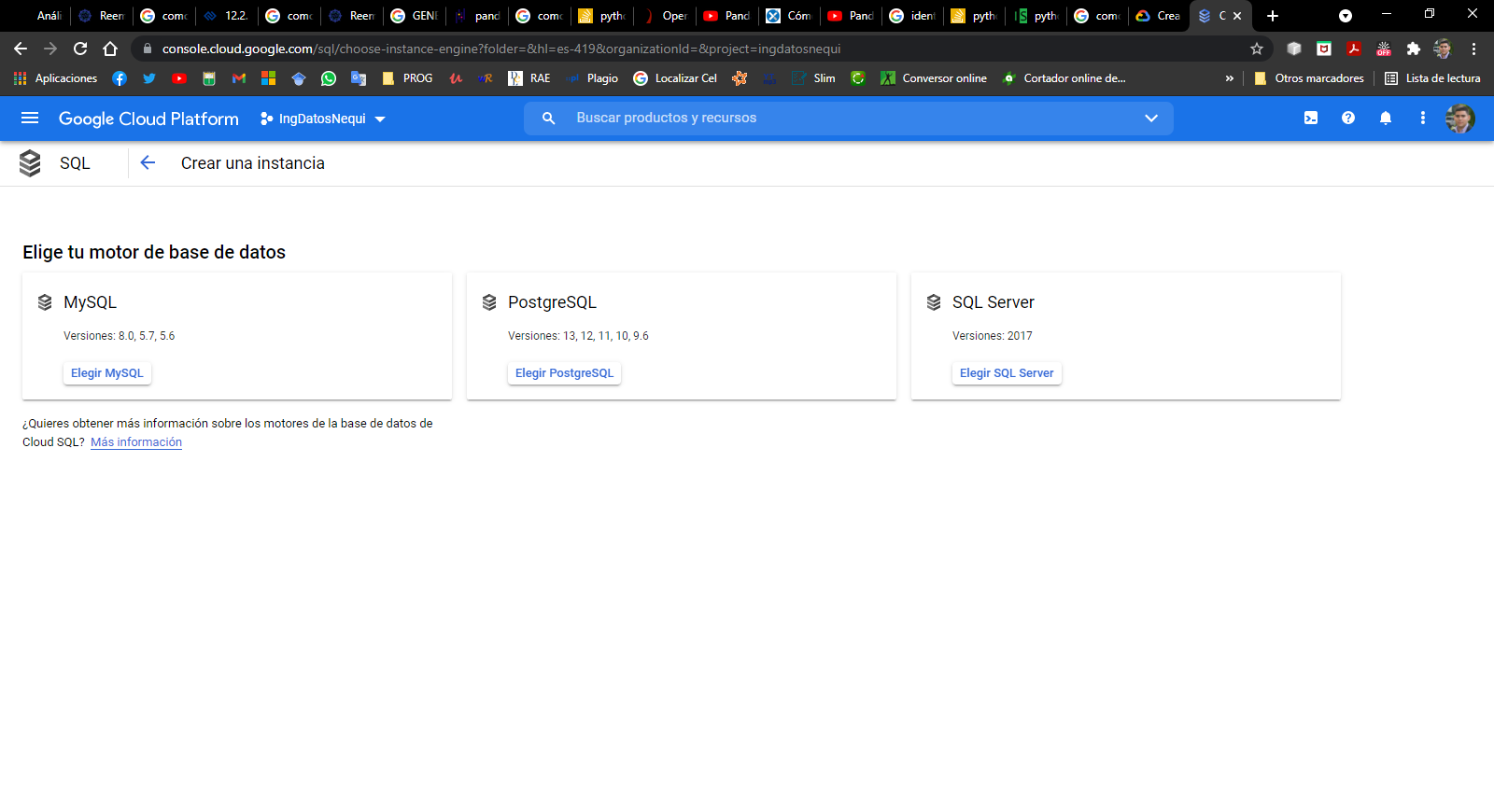
Se identifica que el registro es normal, debido a que es una eps subsidiada de Bogotá, por lo tanto, no se elimina el registro.



**Crear base de datos en SQL**



Se selecciona SQL Server



Paso 3: Definir el modelo de datos

Trazar el modelo de datos conceptuales y explicar por qué se eligió ese modelo.



Se elige este modelo, porque se desea tener un histórico mensual de fallecidos por covid y afiliados por departamento, para así determinar posibles variaciones.

Diseñar la arquitectura y los recursos utilizados.

Indique claramente los motivos de la elección de las herramientas y tecnologías para el

proyecto.

**Proponga con qué frecuencia deben actualizarse los datos y por qué**

Los datos deben actualizarse mensualmente, debido a que el seguimiento a las EPS se realizará cada mes de acuerdo con el número de fallecidos por covid-19.

**Paso 4: Ejecutar la ETL**

Librerias utilizadas

Pandas

Numpy

google.cloud

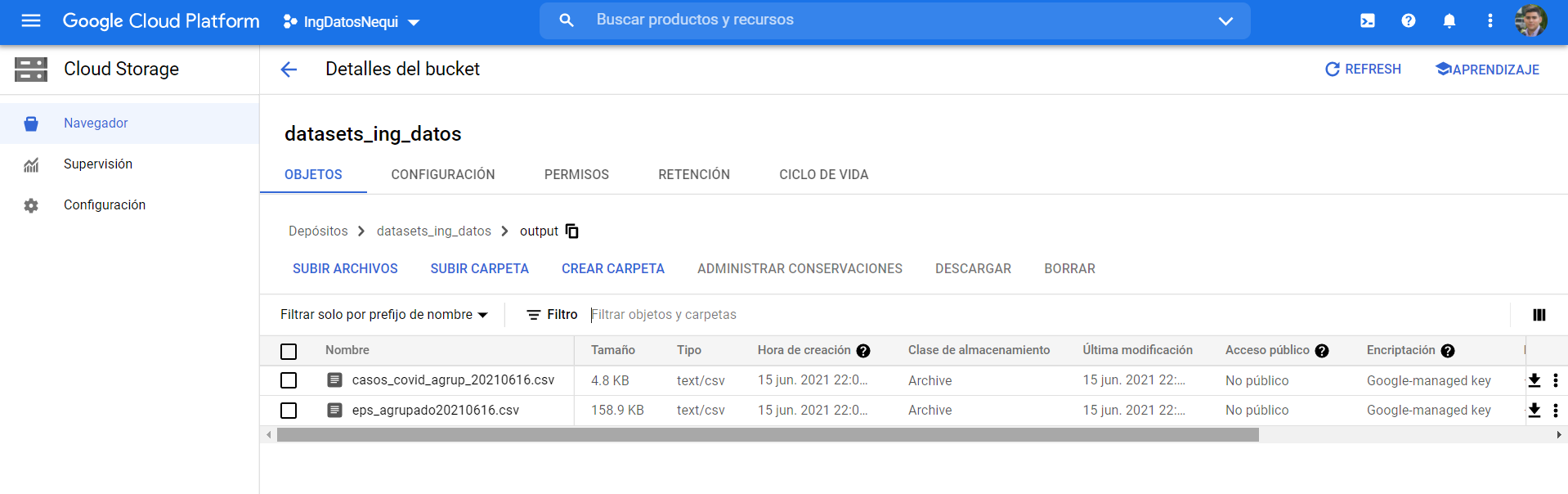
datetime

sodapy

El archivo ingdatosnequi-8799c08e6f65.json contiene le key para acceder al bucket de Google Cloud Storage

Ver archivo prueba\_nequi.py

Resultado final ejecución



**Paso 5: Completar la redacción del proyecto**

**¿Cuál es el objetivo?**

El objetivo del proyecto es determinar la relación existente entre la cantidad de muertos por COVID -19 y la cantidad de afiliados a en cada una de las EPS por departamento y así determinar cuáles EPS se deben intervenir.

**¿Qué preguntas quieres hacer?**

1. ¿Cuál es la relación existente entre los muertos por covid 19 por departamento y cantidad de usuarios afiliados a las EPS?
2. ¿Cuáles son las EPS con más pacientes fallecidos por covid-19 por departamento y a nivel nacional?
3. Realizar análisis de proporciones y determinar ¿cuáles son las EPS con mayor número de fallecidos’

**¿Por qué eligió el modelo que eligió?**

Se elige este modelo, porque se desea tener un histórico mensual de fallecidos por covid y afiliados por departamento, para así determinar posibles variaciones.

Incluya una descripción de cómo abordaría el problema de manera diferente en los

siguientes escenarios:

1. Si los datos se incrementaran en 100x.

Se consumirían a través de APIS y se realizarían opciones de filtrado.

1. Si las tuberías se ejecutaran diariamente en una ventana de tiempo especifica.

3. Si la base de datos necesitara ser accedido por más de 100 usuarios funcionales.

4. Si se requiere hacer analítica en tiempo real, ¿cuales componentes cambiaria a su

arquitectura propuesta?