



**ANÁLISIS DE LAS PRECIPITACIONES EN LOS MUNICIPIOS DE CUNDINAMARCA**

Luis Felipe Vargas Cubillos

Bootcamp análisis de datos

Nivel: Integrador

Grupo: G3

Ejecutor: Jimmy Alexander Muñoz Gutiérrez

6 de Julio de 2024

# OBJETIVOS

* Identificar fuentes de datos relevantes, como bases de datos ambientales, informes gubernamentales, y estudios de investigación.
* Recopilar y limpiar los datos, abordando problemas de calidad, valores faltantes y posibles sesgos.
* Realizar una exploración inicial de los datos para identificar patrones preliminares y posibles áreas de interés.

# IDENTIFICACIÓN DE FUENTE DE DATOS

**Nombre de la fuente de datos:** Precipitaciones Totales Mensuales

**Descripción:** Volumen de precipitaciones mensuales obtenido en la red de estaciones hidrológicas de la Corporación Autónoma Regional de Cundinamarca - CAR.

**Datos suministrados por:** Corporación Autónoma Regional de Cundinamarca (CAR)

**Última Actualización:** 20 de abril de 2024

**Idioma:** español

**Cobertura Geográfica:** Departamental

**Frecuencia de Actualización:** Anual

**Fecha Emisión (aaaa-mm-dd):** 2016-11-01

**Licencia:** Creative Commons Attribution | Share Alike 4.0 International

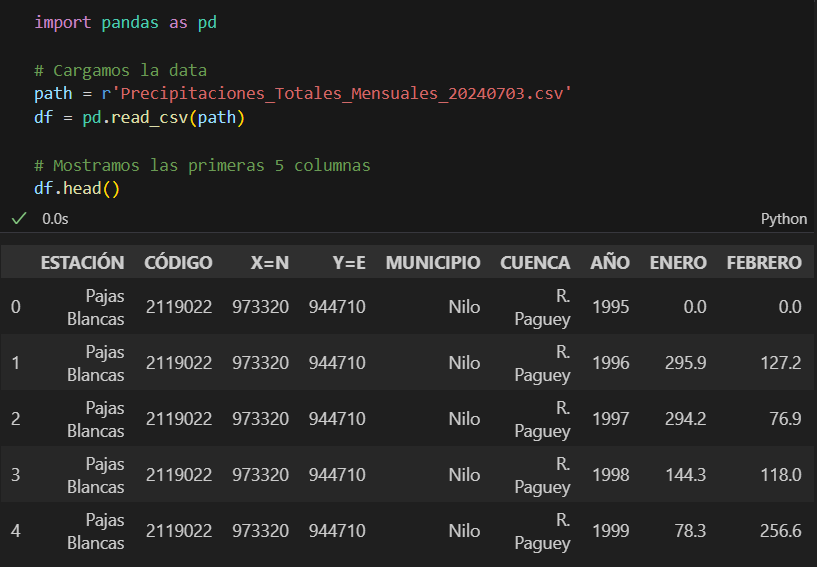
**Enlace de la fuente:** <https://www.datos.gov.co/d/mb4n-6m2g>

**Descripción de los datos**

* Los datos adquiridos nos muestran las precipitaciones (lluvias) mensuales en mm de diferentes municipios de Cundinamarca en un rango de años que va desde 1919 hasta el 2019.
* Se da una georreferencia de la ubicación de las estaciones meteorológicas las cuales están clasificadas por municipio, nombre y código.
* Adicional, el archivo nos muestra la cuenca mas cercana a la que se tomaron los datos de precipitaciones mensuales.

# DOCUMENTACIÓN PROCESO DE LIMPIEZA DE DATOS

1. Se utiliza la biblioteca Pandas de Python para realizar la limpieza y asegurar la consistencia de los datos. El archivo de datos tipo **csv** contiene 6196 filas y 19 columnas. Se cargan los datos crudos en el programa.

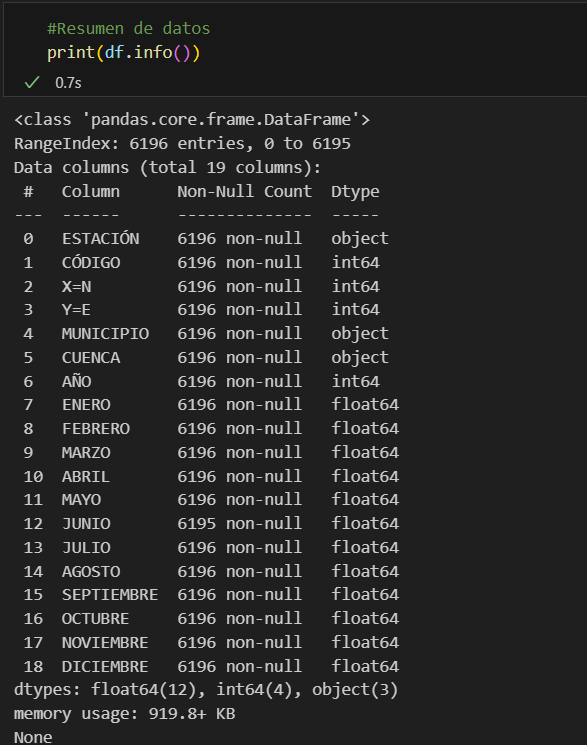


El archivo cargado nos muestra los datos de la precipitación mensual de diferentes municipios de Cundinamarca en un rango de años que va desde 1919 hasta el 2019.

1. Se realiza un resumen de datos con la función **<.info()>**

Con la cual muestra que el archivo contiene:

* 6196 entradas
* 19 columnas
* 12 columnas con datos tipo **float64**
* 4 columnas con datos tipo **int64**
* 3 columnas con datos tipo **object**
* La memoria utilizada son **919.8 KB**

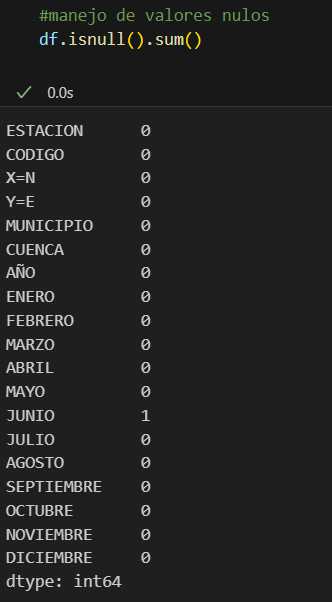


1. Renombrar columnas

Para hacer más rápido el análisis de los datos se renombraron las columnas con acentos del idioma a los mismos nombres, pero sin acento lo cual facilita el manejo de los datos.



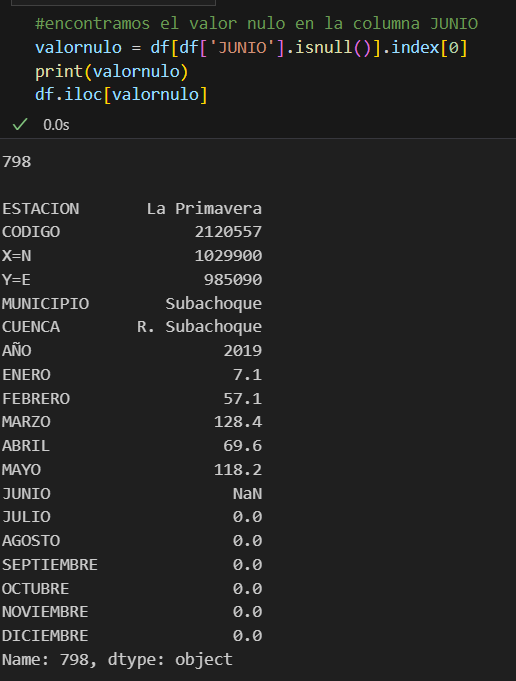
1. Manejo de valores nulos



Se inicia identificando los valores nulos utilizando la función **isnull()**.

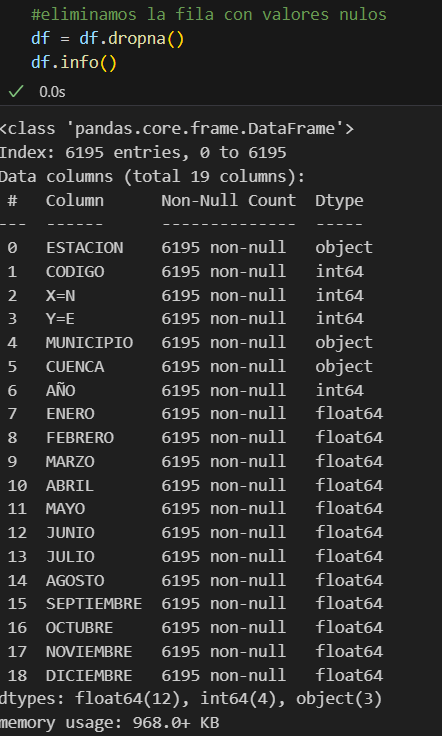
Se muestra que tenemos un valor nulo en la columna “JUNIIO”

Identificamos el valor nulo



Se encuentra que el valor se encuentra en la fila 798 de la estación “La Primavera”

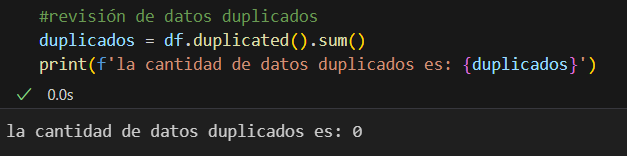
Se eliminan los datos nulos con **dropna(**).



Se verifica que todas lo columnas cuentan ahora con los mismos datos non-null.

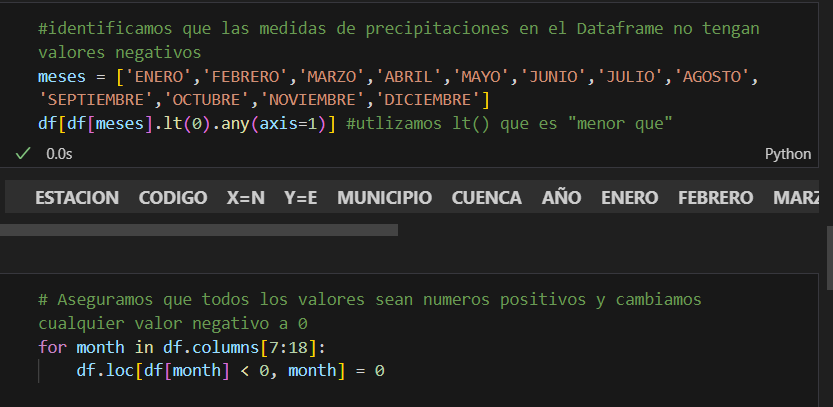
1. Análisis de datos duplicados

Se utiliza la función. duplicated() para identificar filas duplicadas



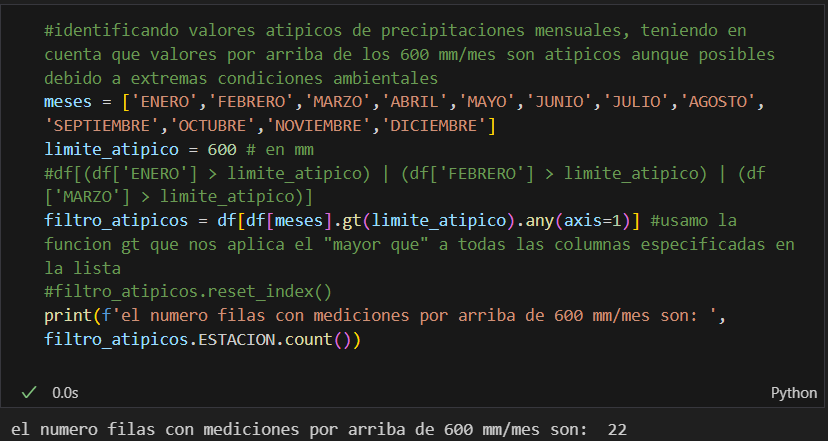
No se encontraron datos duplicados en el dataframe

1. Identificación de valores atípicos

Se asegura que los datos no tengan valores negativos.

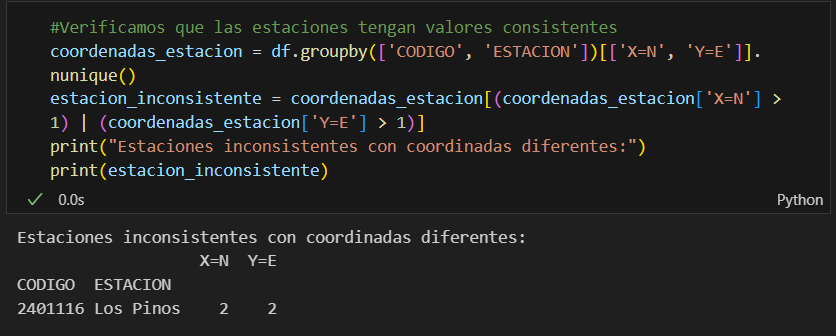
Se utiliza un ciclo for para asegurar la consistencia de los datos.

Verificamos que numero de valores fuera de los rangos normales tenemos, teniendo en cuenta que valores por arriba de los 600 [mm/mes] son atípicos, aunque posibles debido a extremas condiciones ambientales. Por esta razón solo se identifican, pero no se eliminan o modifican.



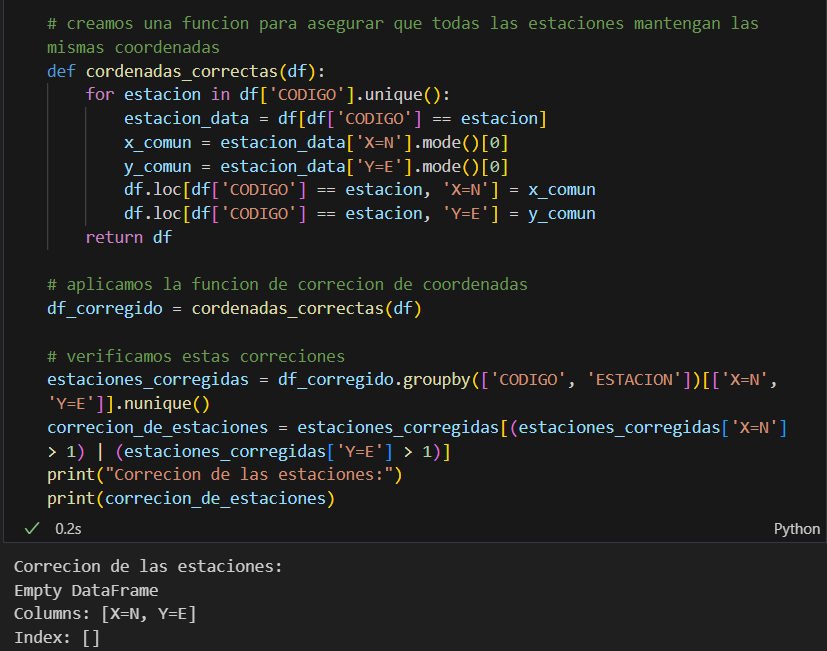
Se encontraron solo 22 filas que tienen valores atípicos por arriba del límite normal especificado.

1. Análisis de inconsistencias en las coordenadas geográficas de las estaciones

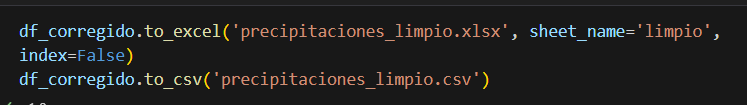


Se encuentra una estación con dos coordenadas diferentes.

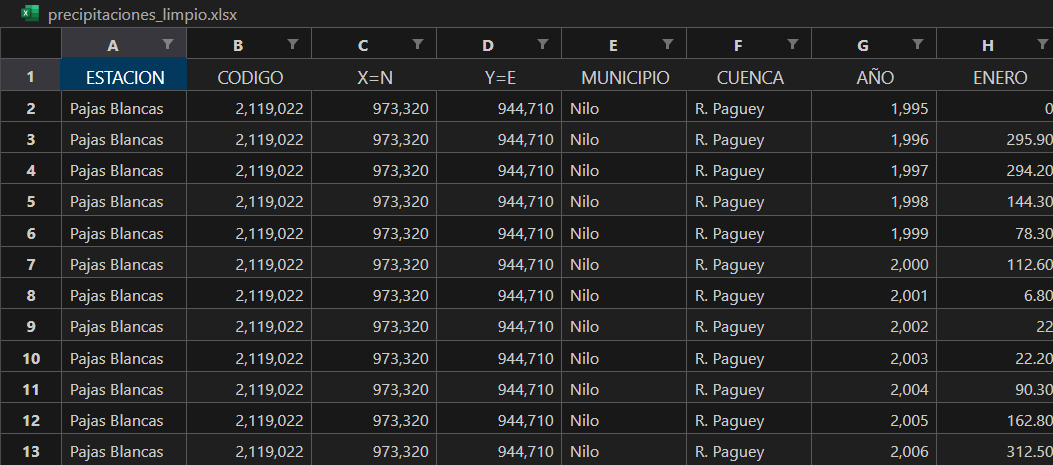
Creamos una función para darle una solo coordenada a los datos de la estación con el nombre Los pinos.



1. Se exportan los datos limpios y consistentes a un archivo Excel y a un archivo CSV.



Archivo Excel



Archivo CSV



# EXPLORACIÓN INICIAL DE LOS DATOS

* 1. Para una exploración inicial de los datos usamos la función describe(). De esta forma obtenemos estadísticas de cada columna con valores numéricos como son:

Cuenta

Media

Desviación estándar

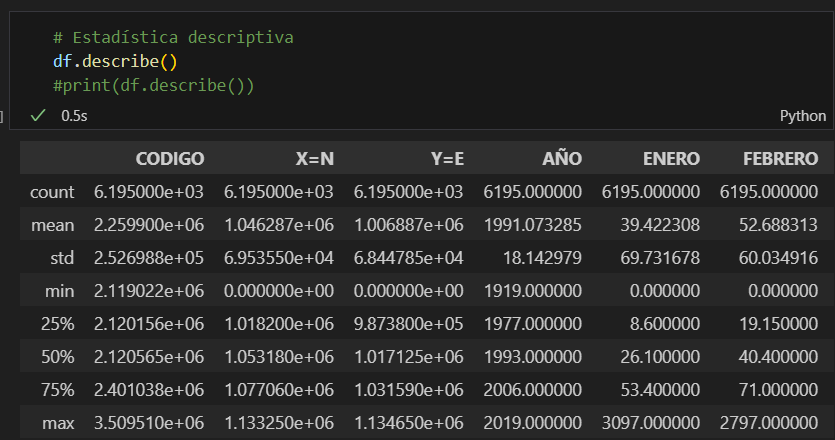
Valor mínimo

Cuartil 25%

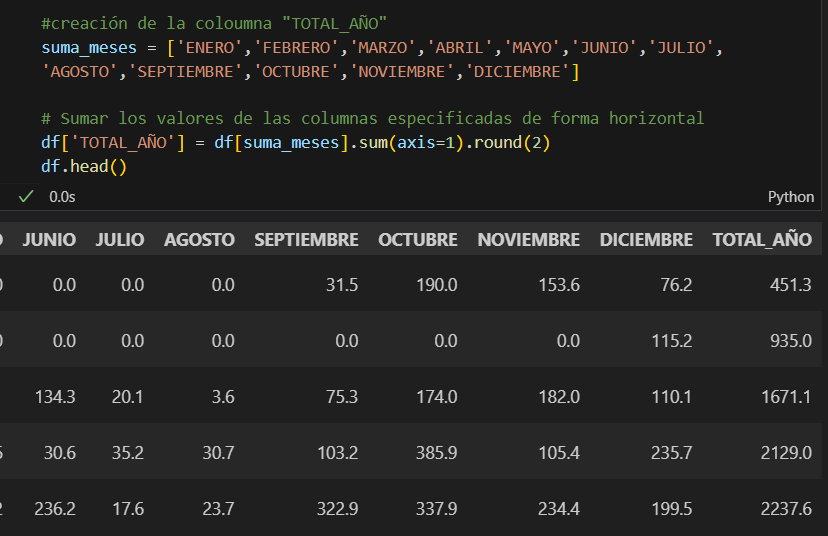
Cuartil 50%

Cuartil 75%

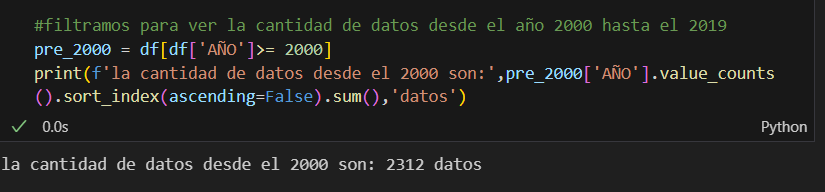
Valor máximo



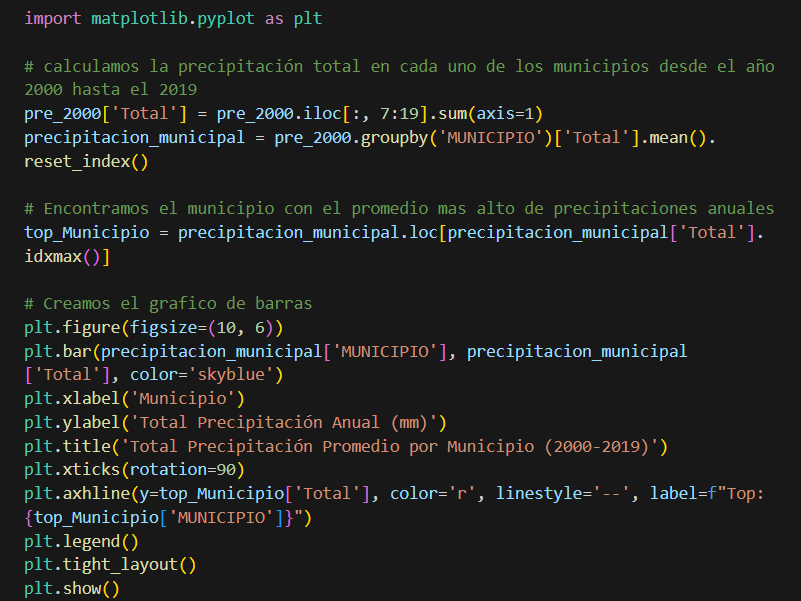
* 1. Añadimos a los datos una columna adicional para totalizar las precipitaciones anuales por cada fila.

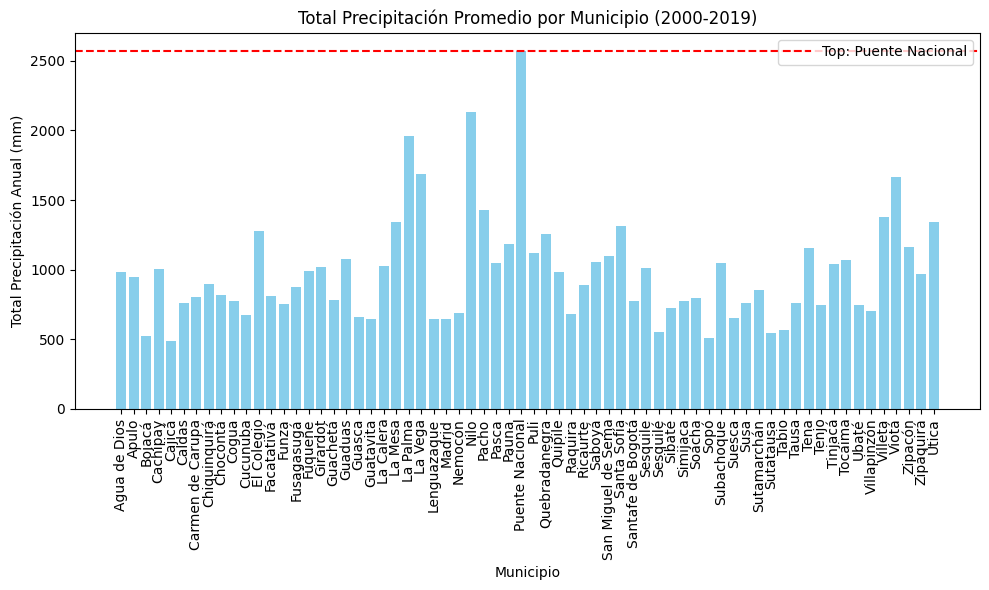


* 1. Se filtran para ver solo los datos de precipitaciones desde el año 2000 al 2019



* 1. Se crea un gráfico con las precipitaciones anuales promedio desde entre el 2000 y 2019 para cada municipio usando los datos del filtro creado anteriormente.





Se encuentra que el municipio que mas lluvias registra en promedio es “Puente Nacional” según los datos analizados en el dataframe.

# CONCLUSIONES

* Se encontró que la data en crudo tenia buenas condiciones para el análisis de datos, para asegurar la consistencia se realizo una limpieza de datos de valores nulos y se utilizaron funciones para asegurar la consistencia de los datos como evitar números negativos, filtrar los valores atípicos y asegurar la consistencia de datos cualitativos como las coordenadas de las estaciones donde se tomaron las muestras de precipitaciones.
* Es necesario tener en cuenta que la data cuenta con 22 filas de valores atípicos que superan los 600 mm de lluvias al mes. Lo que puede hacer varias las mediciones promedio realizadas. Los datos son atípicos, pero no necesariamente están errados ya que es posible obtener estos datos en condiciones extremas climáticas de forma transitoria en algún momento de la medición.