

**Informe 1. Primer avance**

Luisa Fernanda Buriticá & Marcela Echeverri

Grupo de investigación FACOM

18 de marzo de 2022

Universidad de Antioquia

El presente informe tiene como objetivo profundizar en el trabajo realizado por las estudiantes Luisa Fernanda Buriticá y Marcela Echeverri desde el 4 de marzo hasta el día 18 de marzo del 2022.

En primer lugar se generaron 2 bases de datos con extensiones csv para las variables meteorológicas de temperatura en unidades de grados Celsius y precipitación en unidades de milímetros; estas variables tenían información de distintas estaciones meteorológicas ubicadas en varios departamentos y municipios de Colombia, y algunos en diferentes rangos de tiempo; se tiene como primer objetivo caracterizar ambos archivos csv, revisando sus información básica, verificando el orden, los formatos de cada fila y columna, y analizar por medio de las estadísticas básicas a los datos, todo esto principalmente por medio de la librería Pandas en python (Todas las funciones y códigos usados para ejecutar el objetivo se adjuntan en un archivo de Google Colab al final del informe). A continuación se presenta la información básica de ambos archivos:

**Base de datos número 1:**

**Nombre del archivo:** Datos\_Hidrometeorol\_gicos\_Crudos\_-\_Red\_de\_Estaciones\_IDEAM\_\_\_Temperatura.csv

**Variable:** Temperatura

**Unidades:** Celsius [°C]

**Tamaño:** 8,7 G

**Número de columnas:** 12

**Número de filas:** 63’946.972

**Uso de memoria por columna:** 511’575.896 Bytes

El archivo contiene datos de temperatura en ºC, registrados por 544 estaciones meteorológicas alrededor del país, con un muestreo temporal de aproximadamente 12 horas, además, se encuentra información acerca de la zona hidrográfica, longitud, latitud, departamento y municipio. La fecha inicial es el **01 de Enero de 2001** **a las 0:19:05 horas** y la fecha más final para la muestra, es el **13 de Diciembre de 2021 a las 23:55:00 horas**, sin embargo en el IDEAM se siguen actualizando, por lo que en el futuro se buscará registrar y ordenar la información en tiempo real para mantener los datos actualizados.

La siguiente tabla contiene los nombres específicos de cada columna, el tipo de dato (Int, Float, Object, etc...), el nombre de la variable que se usó en python para guardar la información de cada columna y la cantidad de elementos únicos registrados.

**Tabla 1.**

*Descripción de la muestra de temperatura*

| C | Nombre | Tipo | Variable código | Unique |
| --- | --- | --- | --- | --- |
| 0 | CodigoEstacion | int64 | CDE\_HD | 550 |
| 1 | CodigoSensor | int64 | CDS\_HD | 1 |
| 2 | FechaObservacion | Object | FECHA\_HD | 1.859.248,00 |
| 3 | ValorObservado | float64 | - | - |
| 4 | NombreEstacion | Object | NE\_HD | 544 |
| 5 | Departamento | Object | Departamento\_HD | 41 |
| 6 | Municipio | Object | Municipio\_HD | 374 |
| 7 | ZonaHidrografica | Object | - | - |
| 8 | Latitud | float64 | - | - |
| 9 | Longitud | float64 | - | - |
| 10 | DescripcionSensor | Object | - | - |
| 11 | UnidadMedida | Object | - | - |

*C contiene el número de columnas, Nombre, tiene la información del nombre de la columna de la muestra de datos dentro del archivo csv, Tipo, es el tipo de variable que contiene la columna, variable código, es el nombre de la variable dentro del Colab, Unique, es la cantidad de datos individuales que contiene cada columna y la información que aún no identificamos se representa con un guión.*

**Base de datos número 2:**

**Nombre del archivo:**

Precipitaci\_n.csv

**Variable:** Precipitación

**Unidades:** Milímetros [mm]

**Tamaño:** 21,1G

**Número de columnas:** 12

**Número de filas:** 160’665.056

**Uso de memoria por columna:** 160’665054 Bytes

El archivo contiene datos de precipitación en mm, registrados por 772 estaciones meteorológicas alrededor del país, con un muestreo temporal de 10 minutos para unas fechas y para otras el muestreo temporal es de 1 minuto, se pretende generalizar esta información con el fin de tener intervalos de 10 minutos para toda la base de datos, además se encuentra información acerca de la zona hidrográfica, longitud, latitud, departamento y municipio de la toma de la medición. La fecha inicial es el **02 de Enero de 2003 a las 15:20:00 horas** y la fecha final de la muestra de datos es el **13 de Diciembre de 2021 a las 23:59:00 horas**, sin embargo, al igual que con la primera base de datos, en el IDEAM se siguen actualizando, por lo que en el futuro se buscará registrar y ordenar la información en tiempo real para mantener los datos actualizados.

La siguiente tabla contiene los nombres específicos de cada columna, el tipo de dato (Int, Float, Object, etc...), el nombre de la variable que se usó en python para guardar la información de cada columna y la cantidad de elementos únicos registrados.

**Tabla 2.**

*Descripción de la muestra de precipitación*

| C | Nombre | Tipo | Variable código | unique |
| --- | --- | --- | --- | --- |
| 0 | CodigoEstacion | int64 | CDE\_P | 790 |
| 1 | CodigoSensor | int64 | CDS\_P | 1 |
| 2 | FechaObservacion | Object | FECHA\_P | 3.013.294,00 |
| 3 | ValorObservado | float64 | - | - |
| 4 | NombreEstacion | Object | NE\_P | 772 |
| 5 | Departamento | Object | Departamento\_P | 41 |
| 6 | Municipio | Object | Municipio\_P | 469 |
| 7 | ZonaHidrografica | Object | - | - |
| 8 | Latitud | float64 | - | - |
| 9 | Longitud | float64 | - | - |
| 10 | DescripcionSensor | Object | - | - |
| 11 | UnidadMedida | Object | - | - |

*C contiene el número de columnas, Nombre, tiene la información del nombre de la columna de la muestra de datos dentro del archivo csv, Tipo, es el tipo de variable que contiene la columna, variable código, es el nombre de la variable dentro del Colab, Unique, es la cantidad de datos individuales que contiene cada columna, y la información que aún no identificamos se representa con un guión.*

Teniendo la información básica de ambas bases de datos se comenzó con el análisis de las variables y esto es lo que se ha encontrado hasta el momento:

* Se buscó afirmar que no hubiera filas desplazadas a la izquierda o a la derecha (si las hubiera, significa que algún dato, en alguna columna, no está correctamente almacenado o no existe, por lo que se debería revisar en detalle esos casos particulares), se verificó identificando que el tipo de variable por columna fuera igual al tipo de variable dato a dato, y se hizo la suposición de que la primera fila de la base de datos tiene los formatos correctos para hacer una comparación, finalmente hasta ahora las columnas revisadas han cumplido con la suposición inicial.
* Se encontraron varias fechas y horas repetidas, esto se debe a que cada estación toma sus propias mediciones independientes las unas de las otras, por esa razón, en teoría, cada estación debe tener fechas y horas únicas; el muestreo también puede cambiar dependiendo del tipo de sensor que se use.
* Sobre el tiempo, la base de datos de precipitación tiene un muestreo de datos diferente para ciertos intervalos de tiempo, por lo que se busca agrupar los datos por estaciones para luego hacer un acople e identificar las fechas en las que el muestreo cambia, y abrir la posibilidad de que haya otros muestreos diferentes que en un primer momento, no fueron detectados. Luego, se busca ordenar la información de tal forma que todos los datos del archivo tengan el mismo muestreo temporal, identificando la mejor estrategia para llevarlos a un paso de tiempo de 10 min.
* Como se puede observar en la tabla 1 y 2, la columna de los departamentos tiene más elementos únicos que el número de departamentos en Colombia (32 departamentos), esto se debe a que hay datos repetidos que se diferencian con errores de gramática (tildes, mayúsculas, puntos, espacios, etc), la solución es crear listas de chequeo y de revisión que conviertan todas posibilidades en una sola (esperamos que sea retroalimentativo) para guardar todos los errores que encontremos por región y corregirlos de forma automatizada. Esto también aplica para los municipios.

**Anexos:**

**Anexo 1.**

Documento en Google Colab con todas las funciones y códigos (en python 3) diseñados para la revisión de toda la información aquí presentada:

<https://colab.research.google.com/drive/1h7yZzECmTn7kzozc3A4jvxRvGDMnLG7v?usp=sharing>