Fundamentos Tecnologías Emergentes con Python

@SENAComunica

www.sena.edu.co

Limpieza de datos con Python

José Fernando Galindo Suárez
jgalindos@sena.edu.co
2023

Fundamentos Tecnologias Emergentes con Python

PROCESO DE GESTIÓN DE FORMACIÓN PROFESIONAL INTEGRAL

FORMATO GUÍA DE APRENDIZAJE

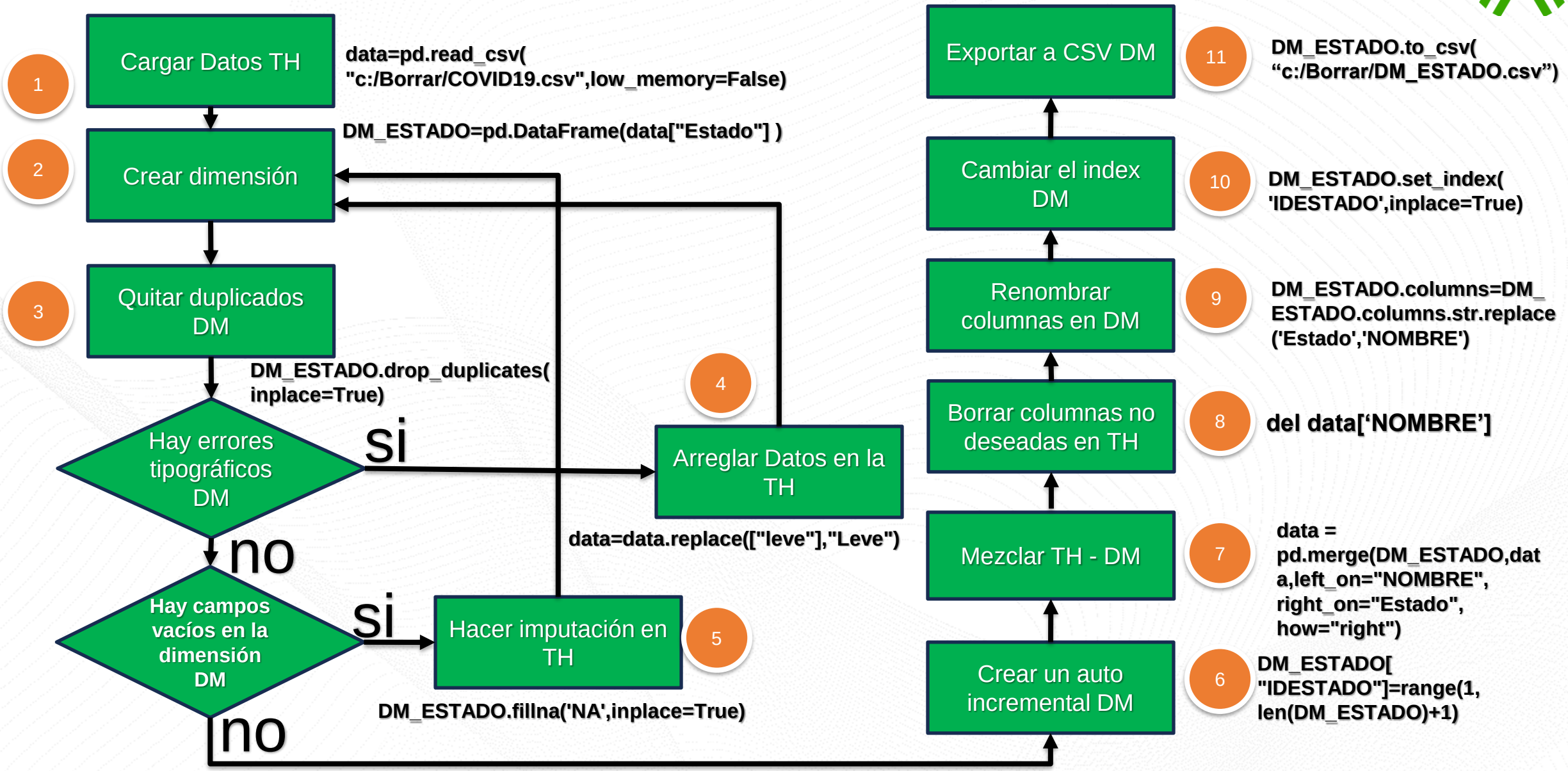
FUNDAMENTOS DE TECNOLOGÍAS EMERGENTES (PYTHON)

IDENTIFICACIÓN DE LA GUIA DE APRENDIZAJE

- Denominación del Programa de Formación: ANALISIS Y DESARROLLO DE SOFTWARE.
- Código del Programa de Formación: 228118
- Nombre del Proyecto: 2605752 DESARROLLO DE SOFTWARE PARA EL CONTROL DE PROCESOS ORIENTADOS A MICROSERVICIOS.
- Fase del Proyecto: EVALUACIÓN
- Actividad de Proyecto: REALIZAR ACTIVIDADES DE VERIFICACIÓN DE CALIDAD DEL SOFTWARE
- Competencia: 220501098 - ADOPCIÓN DE BUENAS PRÁCTICAS EN EL PROCESO DE DESARROLLO DE SOFTWARE.
- Resultados de Aprendizaje Alcanzar:
- 220501098 02 VERIFICAR LA CALIDAD DEL SOFTWARE DE ACUERDO CON LAS PRÁCTICAS ASOCIADAS EN LOS PROCESOS DE DESARROLLO.

- Duración de la Guía: 66 Horas (DIRECTO + INDEPENDIENTE)

Fundamentos Tecnologías Emergentes con Python

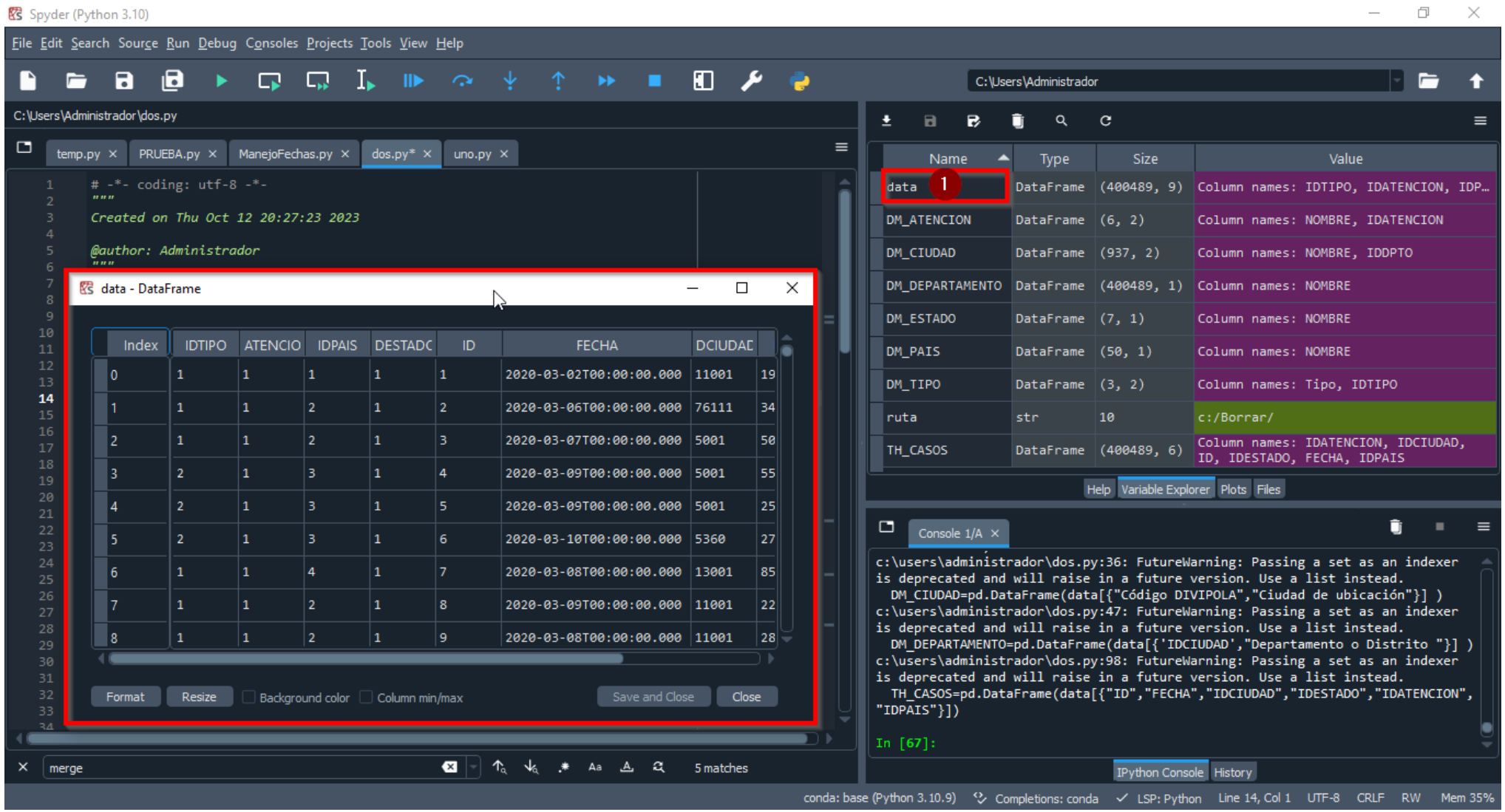


Cargando el dataset

```
import pandas as pd
ruta="c:/Borrar/"
# Cargar los datos
data=pd.read_csv("c:/Borrar/COVID19.csv",low_memory=False)
```

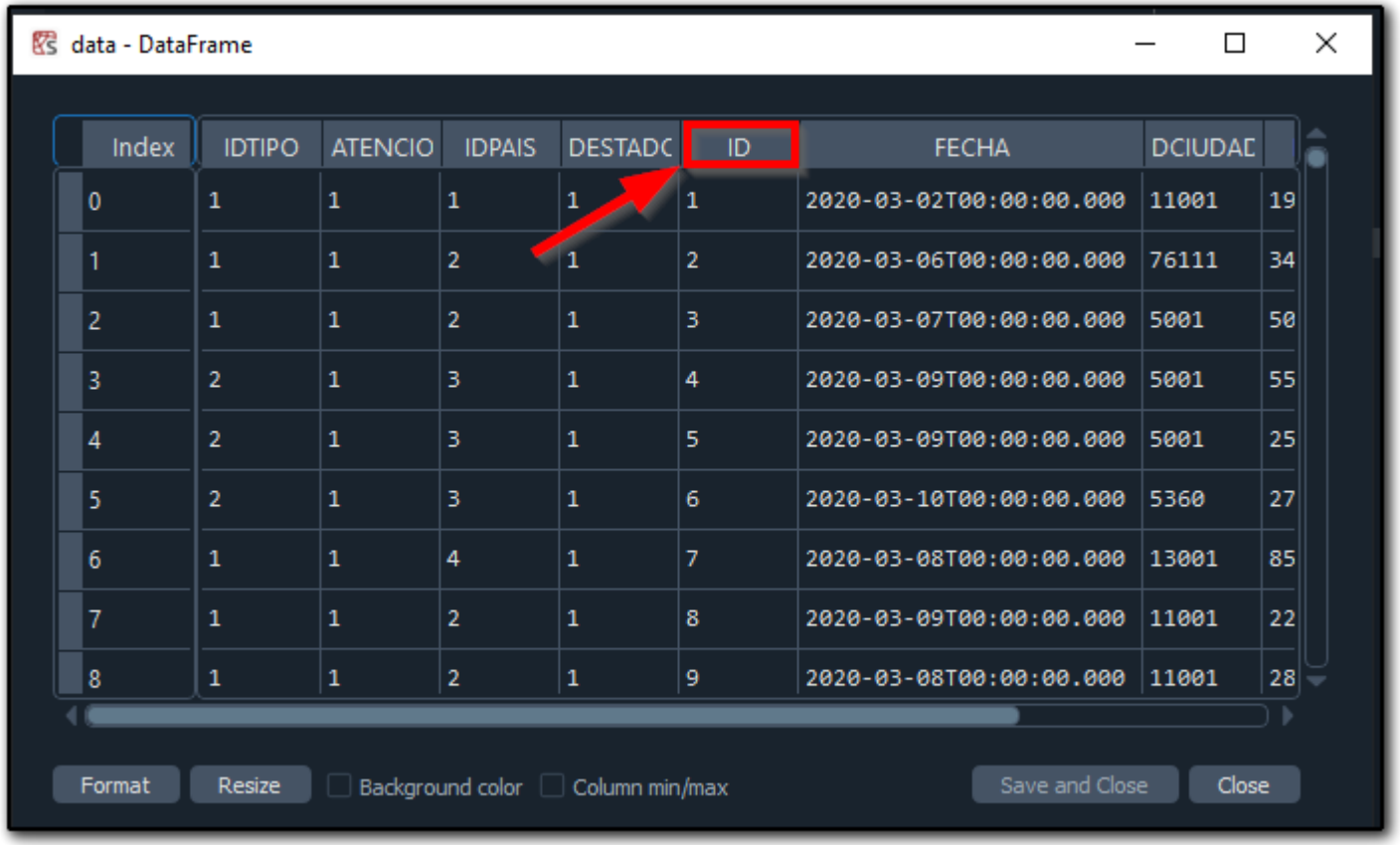
Descargar DATASET **COVID19**

Cargando el dataset



Cambiar el nombre de una columna

data.columns=data.columns.str.replace('ID de caso','ID')



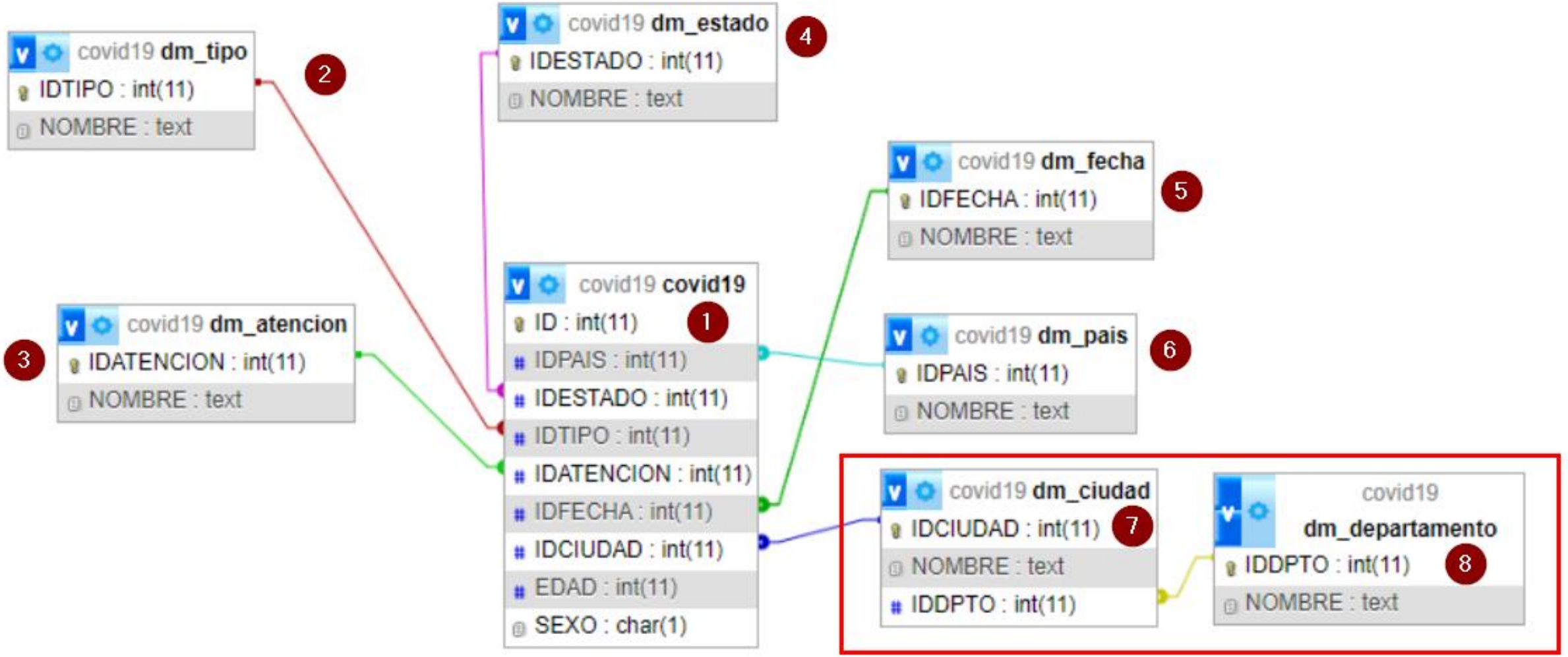
Grabar la dimensión DM_ESTADO en un archivo CSV

```
# Define a 'IDESTADO' como indice
DM_ESTADO.set_index('IDESTADO',inplace=True)
# Guarda a archivo plano
DM_ESTADO.to_csv(ruta+"DM_ESTADO.csv")
```

C: > Borrar > DM_ESTADO.csv

```
1   IDESTADO,NOMBRE
2   1,Leve
3   2,Asintomático
4   3,Fallecido
5   4,NA
6   5,Moderado
7   6,Grave
```

Modelo Estrella y Copo de Nieve



Modelo Fisico de la Dimension y carga de datos

1
```
CREATE TABLE DM_TIPO(
IDTIPO INTEGER PRIMARY KEY AUTO_INCREMENT,
NOMBRE VARCHAR(30)
);
```

2
```
ALTER TABLE COVID19 ADD CONSTRAINT TIPOFK FOREIGN
KEY(IDTIPO)
REFERENCES DM_TIPO(IDTIPO);
```

3
```
LOAD DATA INFILE 'c:/Borrar/DM_TIPO.csv' INTO TABLE DM_TIPO
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 1 LINES
(IDTIPO,NOMBRE);
```

VALIDAR EL MODELO DE BASE DE DATOS

```
MariaDB [covid19]> SELECT * FROM VCOVID19 ;
+-----------------+--------+
| DM_ATENCION     | FILAS  |
+-----------------+--------+
| DM_ATENCION     |      6 |
| DM_CIUDAD       |    937 |
| DM_DEPARTAMENTO |     33 |
| DM_ESTADO       |      6 |
| DM_PAIS         |     44 |
| DM_TIPO         |      3 |
| DM_FECHA        |    149 |
| TH_COVID19      | 400489 |
+-----------------+--------+
8 rows in set (0.185 sec)
```

```
CREATE OR REPLACE VIEW VCOVID19 AS
SELECT 'DM_ATENCION',COUNT(*) FILAS FROM DM_ATENCION
UNION
SELECT 'DM_CIUDAD',COUNT(*) FILAS FROM DM_CIUDAD
UNION
SELECT 'DM_DEPARTAMENTO',COUNT(*) FILAS FROM DM_DEPARTAMENTO
UNION
SELECT 'DM_ESTADO',COUNT(*) FILAS FROM DM_ESTADO
UNION
SELECT 'DM_PAIS',COUNT(*) FILAS FROM DM_PAIS
UNION
SELECT 'DM_TIPO',COUNT(*) FILAS FROM DM_TIPO
UNION
SELECT 'DM_FECHA',COUNT(*) FILAS FROM DM_FECHA
UNION
SELECT 'TH_COVID19',COUNT(*) FILAS FROM COVID19
;
```

VALIDAR CONSTRAINT EN LA BASE DE DATOS

```
CREATE OR REPLACE VIEW ESQUEMA AS
SELECT
CONSTRAINT_NAME,TABLE_SCHEMA,TABLE_NAME,CONSTRAINT_TYPE
FROM INFORMATION_SCHEMA.TABLE_CONSTRAINTS
WHERE CONSTRAINT_SCHEMA='COVID19';
```

```
MariaDB [covid19]> SELECT * FROM ESQUEMA;
+-----------------+--------------+-----------------+-----------------+
| CONSTRAINT_NAME | TABLE_SCHEMA | TABLE_NAME      | CONSTRAINT_TYPE |
+-----------------+--------------+-----------------+-----------------+
| PRIMARY         | covid19      | covid19         | PRIMARY KEY     |
| ATENCIONFK      | covid19      | covid19         | FOREIGN KEY     |
| CIUDADFK        | covid19      | covid19         | FOREIGN KEY     |
| ESTADOFK        | covid19      | covid19         | FOREIGN KEY     |
| PAISFK          | covid19      | covid19         | FOREIGN KEY     |
| TIPOFK          | covid19      | covid19         | FOREIGN KEY     |
| PRIMARY         | covid19      | dm_atencion     | PRIMARY KEY     |
| PRIMARY         | covid19      | dm_ciudad       | PRIMARY KEY     |
| DPTOFK          | covid19      | dm_ciudad       | FOREIGN KEY     |
| PRIMARY         | covid19      | dm_departamento | PRIMARY KEY     |
| PRIMARY         | covid19      | dm_estado       | PRIMARY KEY     |
| PRIMARY         | covid19      | dm_fecha        | PRIMARY KEY     |
| PRIMARY         | covid19      | dm_pais         | PRIMARY KEY     |
| PRIMARY         | covid19      | dm_tipo         | PRIMARY KEY     |
+-----------------+--------------+-----------------+-----------------+
14 rows in set (0.043 sec)
```

RETO EVIDENCIA

DESARROLLE LAS SIGUIENTES ACTIVIDADES:

- **LA DIMENSIONES CON PYTHON DE ACUERDO AL MODELO DE DATOS**
- **EXPORTE LOS DATOS DE CADA DIMENSION A UN ARCHIVO PLANO**
- **CONSTRUYE LA BASE DE DATOS COVID19 EN MYSQL**
- **CONSTRUYA LAS TABLAS DE CADA DIMENSION Y LA TABLA DE HECHO EN MYSQL**
- **CARGUE LOS DATOS DE CADA DIMENSION Y LA TABLA DE HECHO EN MYSQL**

ENTREGUE AL INSTRUCTOR EL SCRIPT DE PYTHON, SQL Y LA GRAFICA DEL MODELO DE DATOS DESDE PHPMYADMIN EN UN ARCHIVO COMPRIMIDO LLAMADO “EVIDENCIA1.zip”

G R A C I A S

Línea de atención al ciudadano: 01 8000 910270
Línea de atención al empresario: 01 8000 910682

@SENAComunica

www.sena.edu.co

https://cienciadedatos.net/machine-learning-python.html