

DATA PROJECT 1: EQUIPO 2

FAN, FRAN, DARÍO Y MIGUEL

CONTENIDOS

- 
- 01 Introducción
 - 02 ETL calidad vida Valencia
 - 03 ETL clientes potenciales
 - 04 BD y calidad del dato
 - 05 Mejora y Visualización
 - 06 Conclusiones

CONTENIDOS

- 
- 01 Introducción
 - 02 ETL calidad vida Valencia
 - 03 ETL clientes potenciales
 - 04 BD y calidad del dato
 - 05 Mejora y Visualización
 - 06 Conclusiones

EQUIPO



MIGUEL MORATILLA

DATA ENGINEERING
PHD AEROSPACE ENG.



FRANCISCO ROSILLO

DATA QUALITY & ENGINEERING
ING. TELECOMUNICACIONES



DARÍO FERNÁNDEZ

DATA ANALYTICS
ECONOMISTA



FAN WU

BUSINESS INTELLIGENCE
ADE

PLANIFICACIÓN DE TRABAJO

Data Project 1 Visible para el Espacio de trabajo Tablero

To do + Añada una tarjeta

Doing

- Concepto dashboard empresa
- Construir Tableau
- Ppt
- Video
- Limpiar tablas
- Tabla Casas
- From datos to SQL
- Ponderaciones según encuesta
- Convertir .ipynb a modulos Python
- Funcion intersección y puntos python

Done

- NiFi para extracción D 1 M
- Data Catalogue
- Base dato casas D 1 M
- Calidad del dato Jupyter
- Python puntuación cliente DF F
- Limpieza de datos M
- Tabla Casas
- From datos to SQL M
- Ponderaciones según encuesta DF
- Convertir .ipynb a modulos Python D 1 DF M
- Funcion intersección y puntos python F M

Herramientas a usar

- + Añada otra lista
- Docker
- Python
- PostgreSQL
- Tableau
- Jupyter
- NiFi

INTRODUCCIÓN

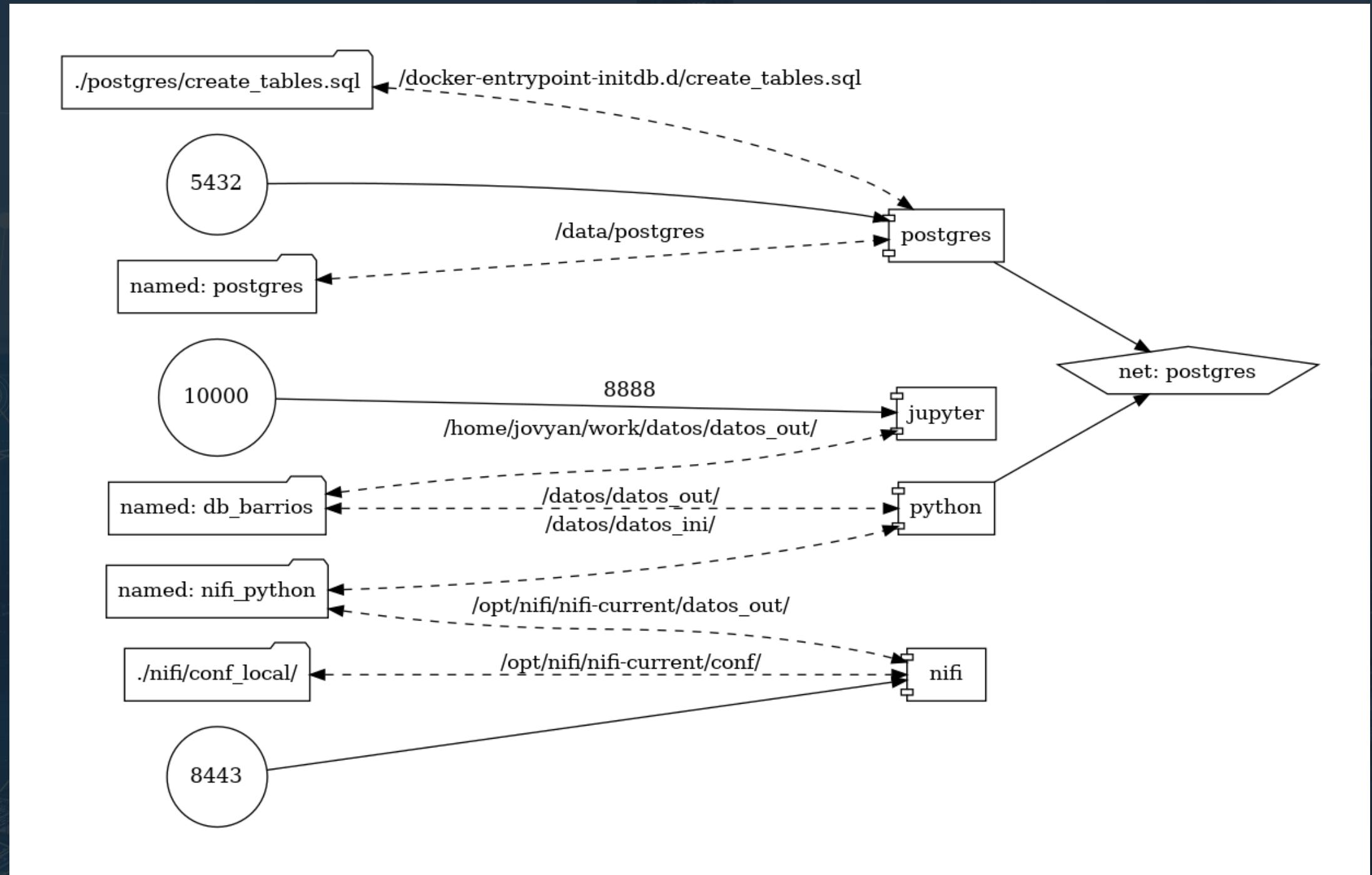
Propuesta: ofrecer un mapa de calidad de vida individualizado por cliente para la ciudad de Valencia

The screenshot shows the official website of the City of Valencia's Open Data portal. At the top, there is a banner featuring the City Hall building and the text "Ayuntamiento" and "Datos abiertos". Below the banner, there is a navigation bar with links to "DATOS ABIERTOS", "CATÁLOGO DE DATOS", "CUADRO DE INDICADORES", "API'S DISPONIBLES", and "APLICACIONES". A secondary navigation bar at the bottom includes links to "AYUNTAMIENTO", "GOBIERNO ABIERTO", and "DATOS ABIERTOS". The main content area features a search bar labeled "Buscar un conjunto de datos" and a row of icons representing various data categories such as transport, health, buildings, cameras, etc. Below this, there are two sections: "Estadísticas de los datasets" (Statistics of datasets) showing "Número de Datasets" (213) and "Descargas" (44.347), and "¿Qué son?" (What are they?) which provides a brief explanation of Open Data principles.

The screenshot shows a survey form titled "Encuesta sobre la compra o alquiler de vivienda". The introduction states: "Esta encuesta sirve para trazar patrones de compra que nos ayuden a entender qué tipo de vivienda busca una persona." The form includes an email input field with "mimove14@gmail.com (no compartidos)" and a "Cambiar de cuenta" link. The survey consists of two questions with multiple choice answers. The first question is "¿Qué edad tienes?" with options: "Entre 20 y 24 años", "Entre 25 y 39 años", "Entre 40 y 60 años", and "Más de 60 años". The second question is "¿Tienes hijos?" with options: "No", "Sí, solo un hijo", and "Sí, tengo más de un hijo".

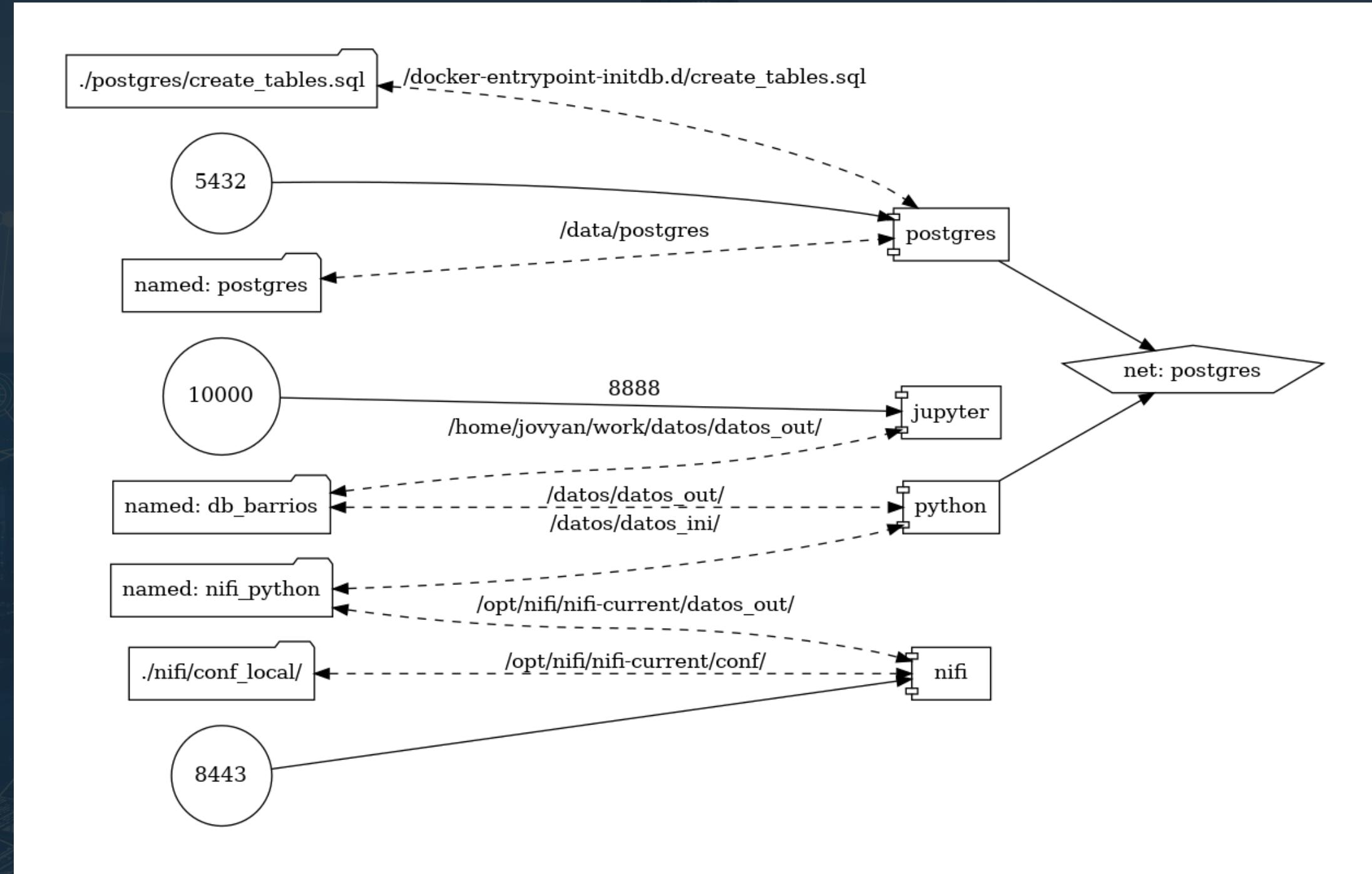
DISEÑO DE ARQUITECTURA

<https://github.com/pmsipilot/docker-compose-viz>



DISEÑO DE ARQUITECTURA

```
docker run --user=root --rm -it --name dcv -v $PWD:/input pmsipilot/docker-compose-viz render -rm image docker-compose.yml
```



CONTENIDOS

- 01 Introducción
- 02 ETL calidad vida Valencia
- 03 ETL clientes potenciales
- 04 BD y calidad del dato
- 05 Mejora y Visualización
- 06 Conclusiones



ETL CALIDAD VIDA VALENCIA

Indicadores seleccionados

- Zonas Verdes
- Distribución de Hospitales
- Distribución de colegios
- Nivel de ruido
- Limpieza
- Puntos de recarga de coches eléctricos
- Transporte público

ETL CALIDAD VIDA VALENCIA

Problemática con los datos

- Inconsistencia en los datos Ayuntamiento de Valencia
- Representación por barrios con polígonos que los delimiten
- Información de los datos en diferentes formas (polígonos, puntos, zonas, etc).

Extracción



Transformación



Carga



ETL CALIDAD VIDA VALENCIA

Problemática con los datos

- Inconsistencia en los datos Ayuntamiento de Valencia
- Representación por barrios con polígonos que los delimiten
- Información de los datos en diferentes formas (polígonos, puntos, zonas, etc).

Extracción



Transformación

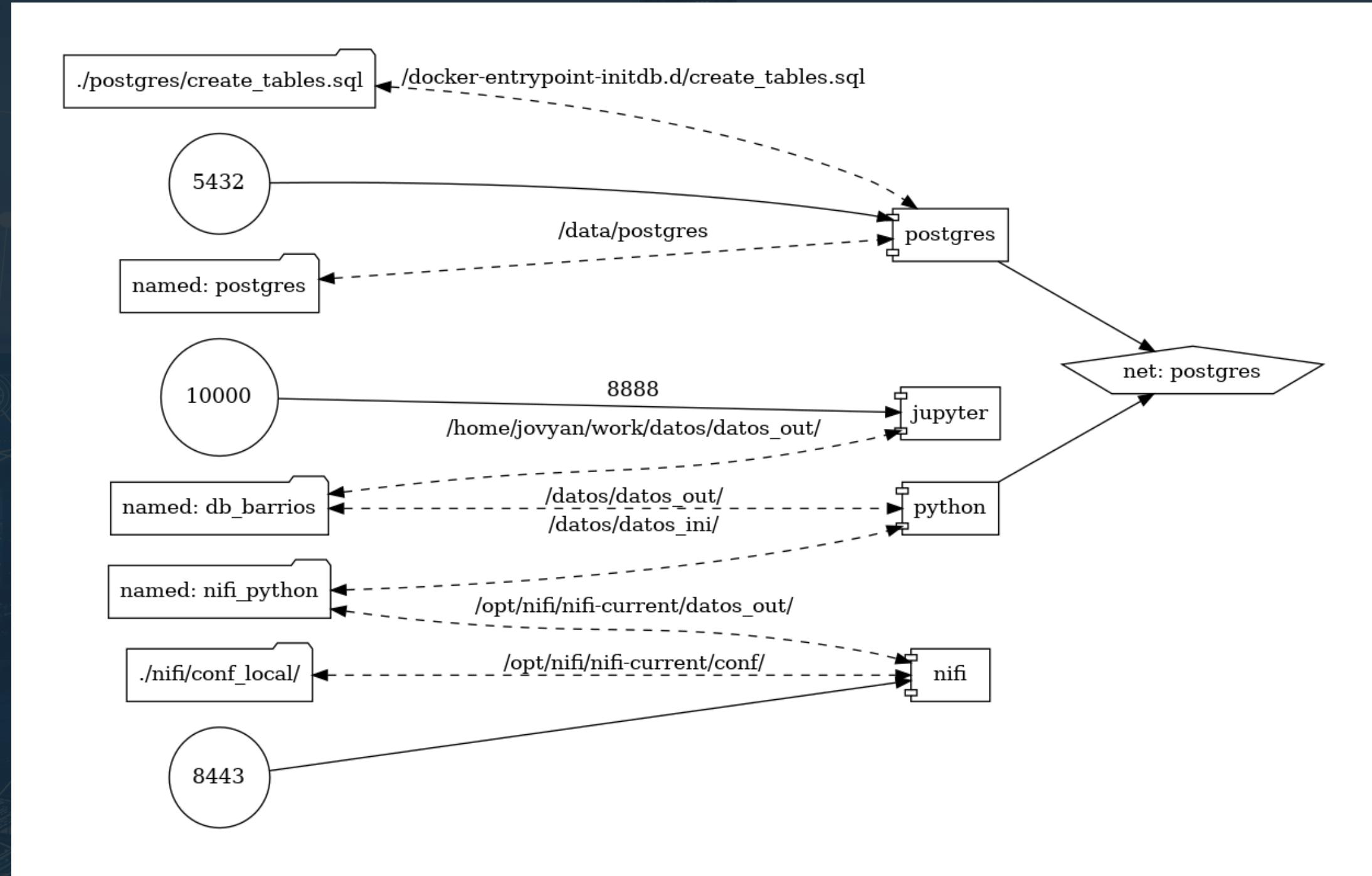


Carga



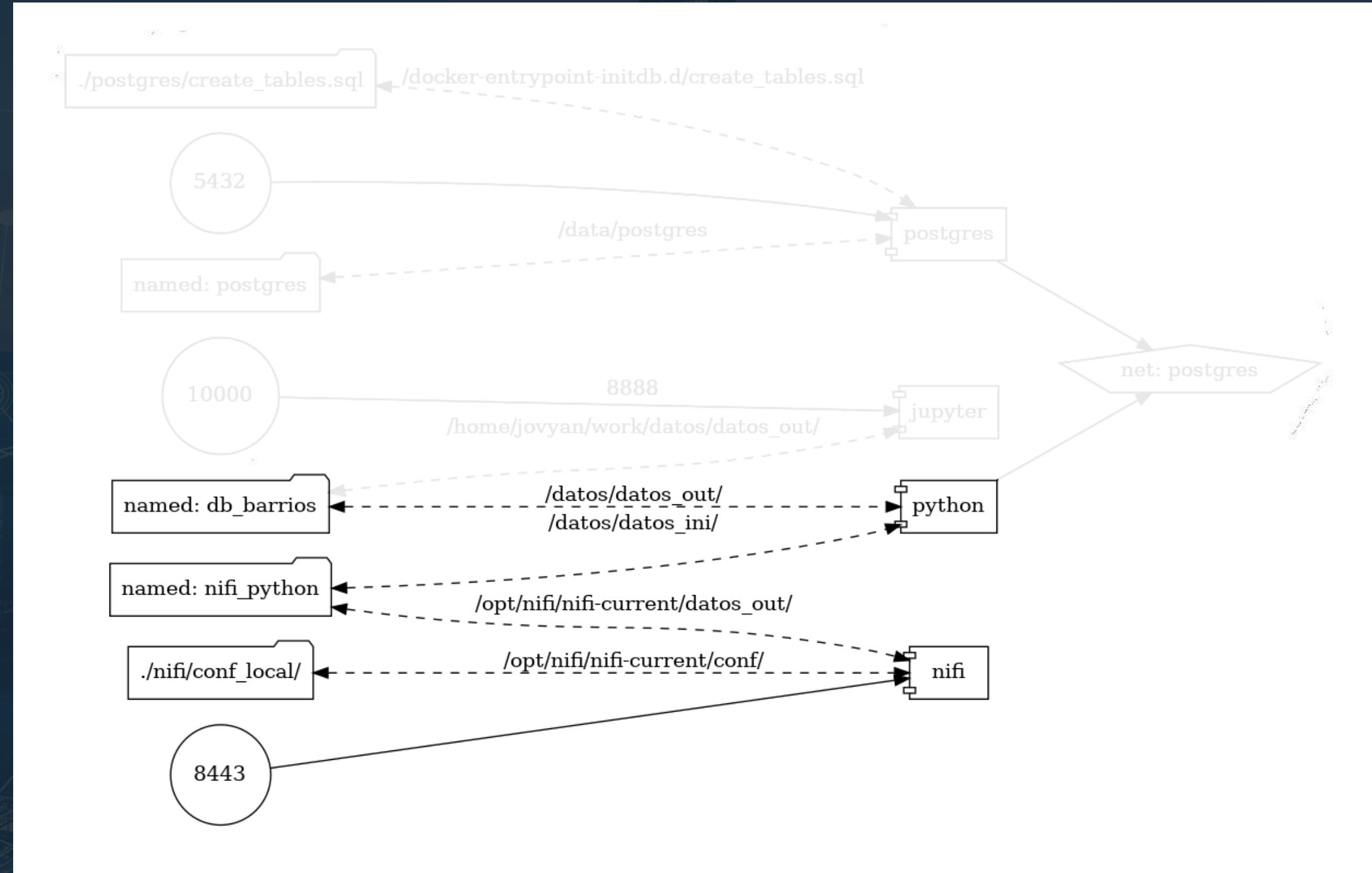
DISEÑO DE ARQUITECTURA

```
docker run --user=root --rm -it --name dcv -v $PWD:/input pmsipilot/docker-compose-viz render -rm image docker-compose.yml
```



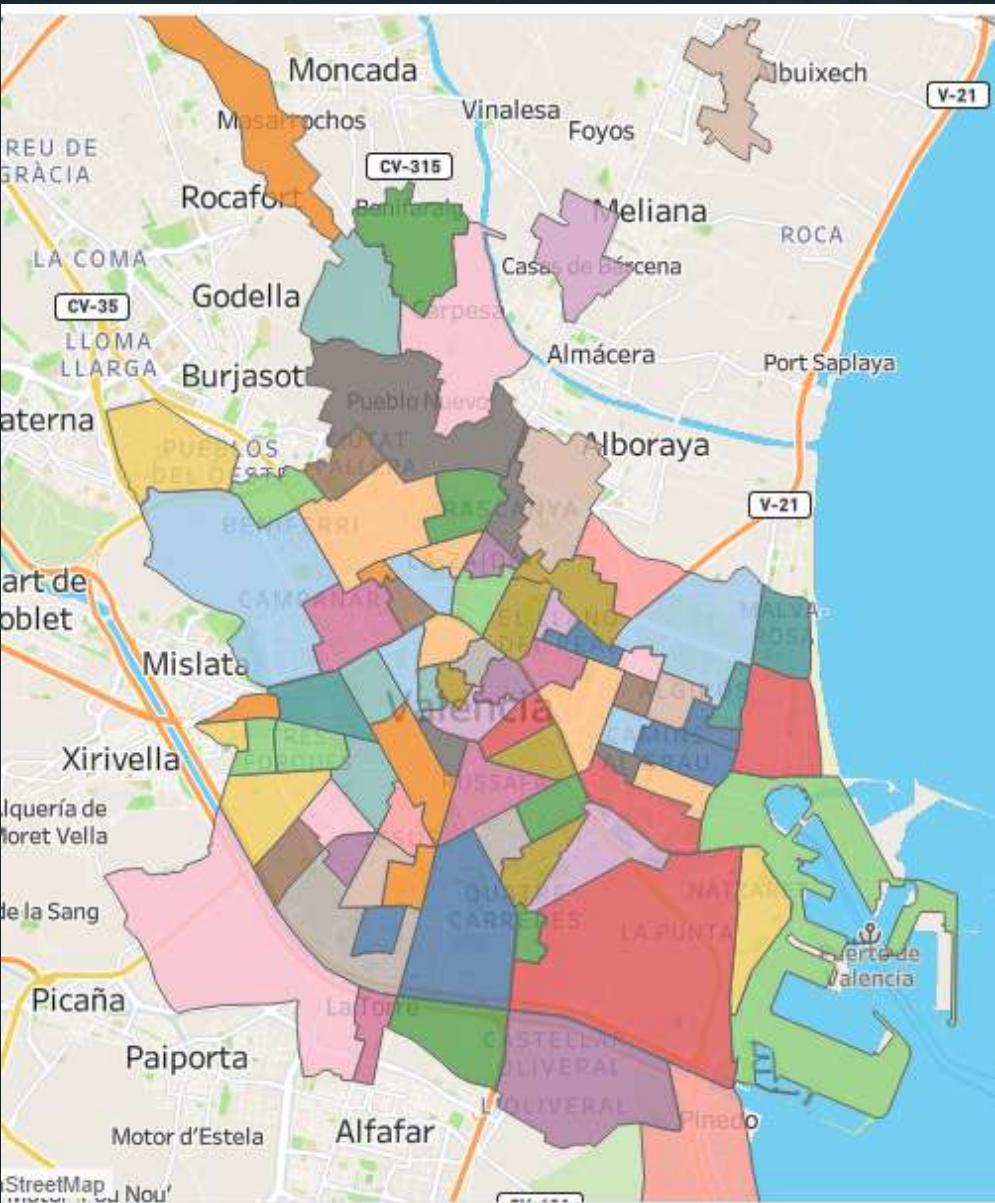
DISEÑO DE ARQUITECTURA

```
docker run --user=root --rm -it --name dcv -v $PWD:/input pmsipilot/docker-compose-viz render -rm image docker-compose.yml
```

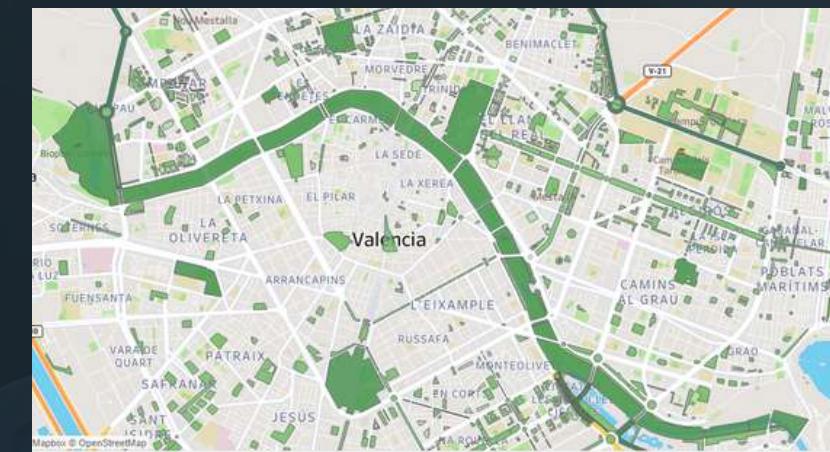


ETL CALIDAD VIDA VALENCIA

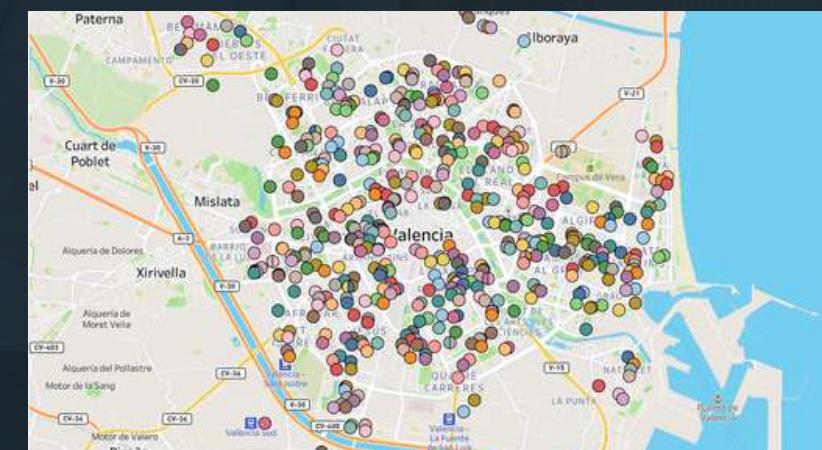
Distribución barrios



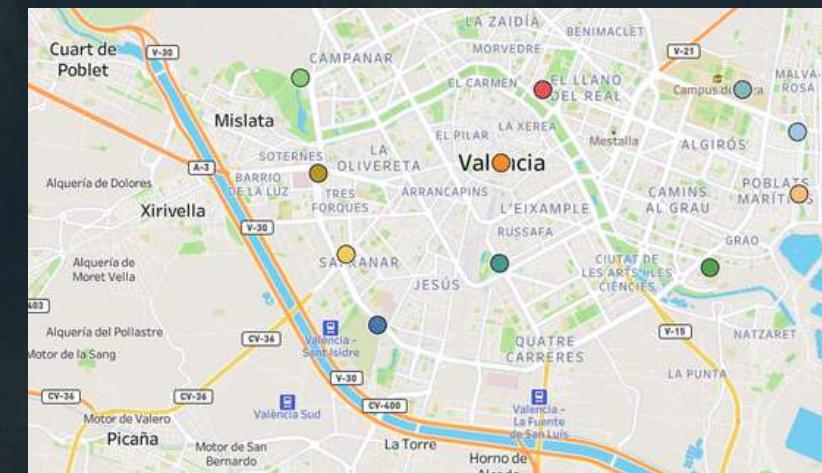
Zonas Verdes



Centros Educativos



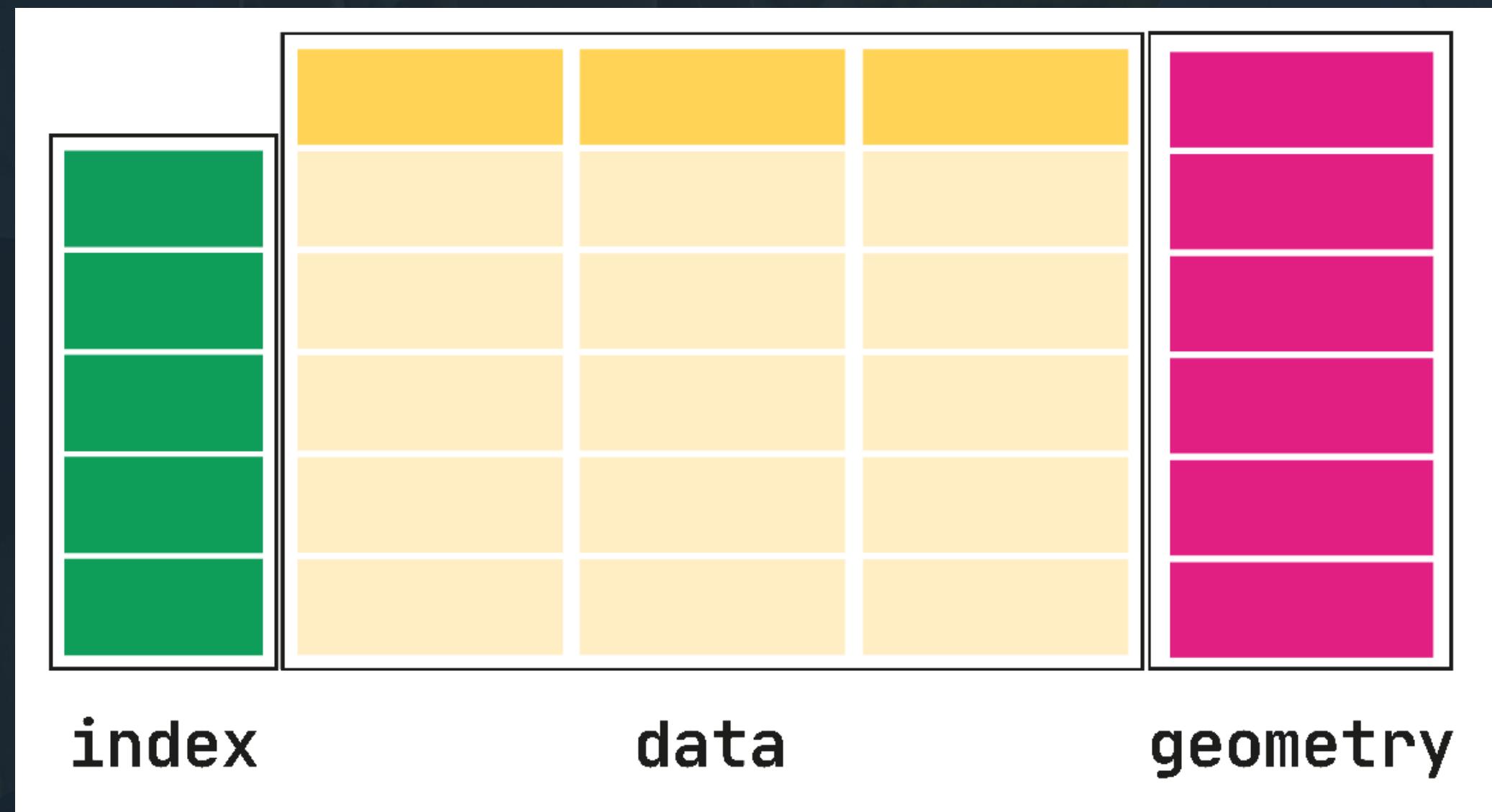
Estaciones contaminación



ETL CALIDAD VIDA VALENCIA

USO DE GEOPANDAS

- Librería basada en Pandas que permite trabajar con dataframes geoespaciales



ETL CALIDAD VIDA VALENCIA

INTERSECCIÓN BARRIOS-POLÍGONOS

```
1 import geopandas as gpd
2
3 def interseccion_poligonos(barrios_in, file2_json):
4
5     file2 = gpd.GeoDataFrame.from_features(file2_json)
6     file2.crs = 'epsg:4326' #GPS
7
8     merged = gpd.overlay(barrios_in, file2, how = 'intersection')
9
10    ## Código adicional para pasar de GPS a Cylindrical equal-area projection
11
12    merged['areaZonaVerde'] = merged.geometry.area
13
14    merged_areas = merged.groupby('nombre_barrio')['areaVariable'].sum()
15
16    barrios_out = barrios_in.merge(merged_areas, on='nombre_barrio', how='left')
17
18    barrios_out['%zona_verde'] = barrios_out.geometry.area/barrios_out['areaZonaVerde']
19
20    return barrios_out
```

ETL CALIDAD VIDA VALENCIA

EXTRAPOLACIÓN BARRIOS-CONTAMINACIÓN

```
import geopandas as gpd

def interpolacion_puntos(barrios_in, file2_json):
    file2 = gpd.GeoDataFrame.from_features(file2_json)
    file2.crs = 'epsg:4326' #GPS

    ## Código adicional para pasar de GPS a Cylindrical equal-area projection

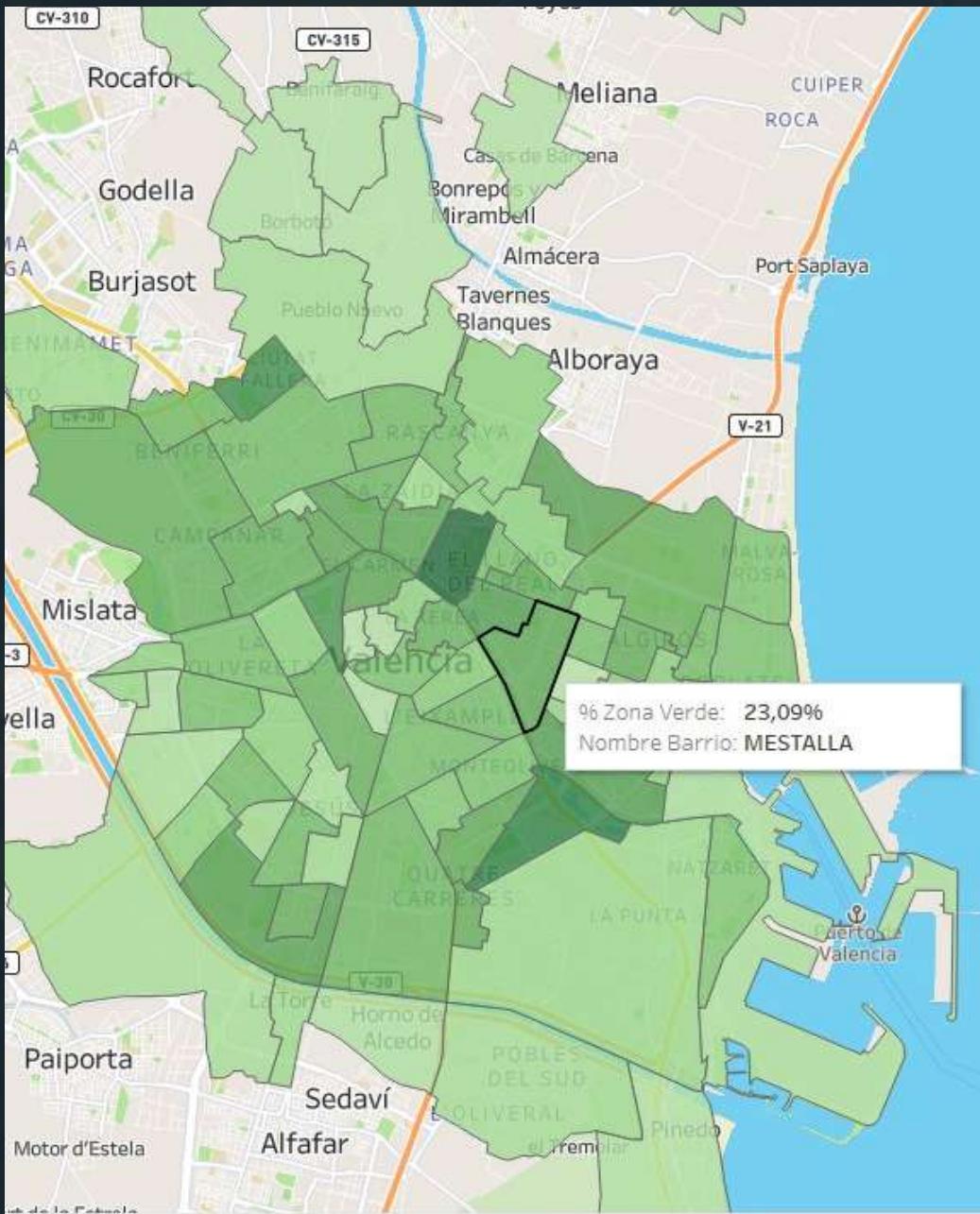
    barrios_out= gpd.sjoin_nearest(barrios_in,file2, how="left", max_distance=1)

    barrios_out = barrios_out[~barrios_out['nombre_barrio'].duplicated()]

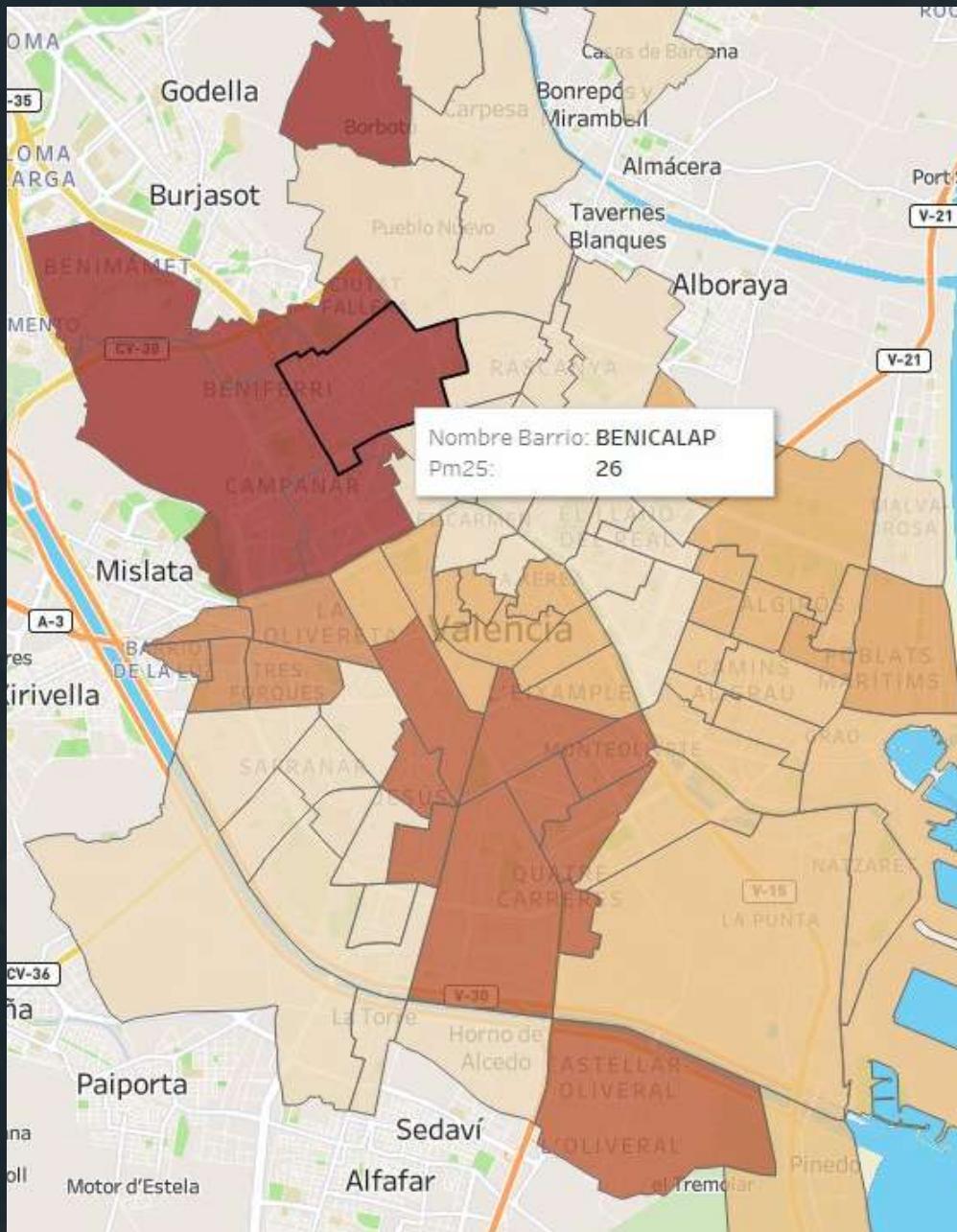
    return barrios_out
```

ETL CALIDAD VIDA VALENCIA

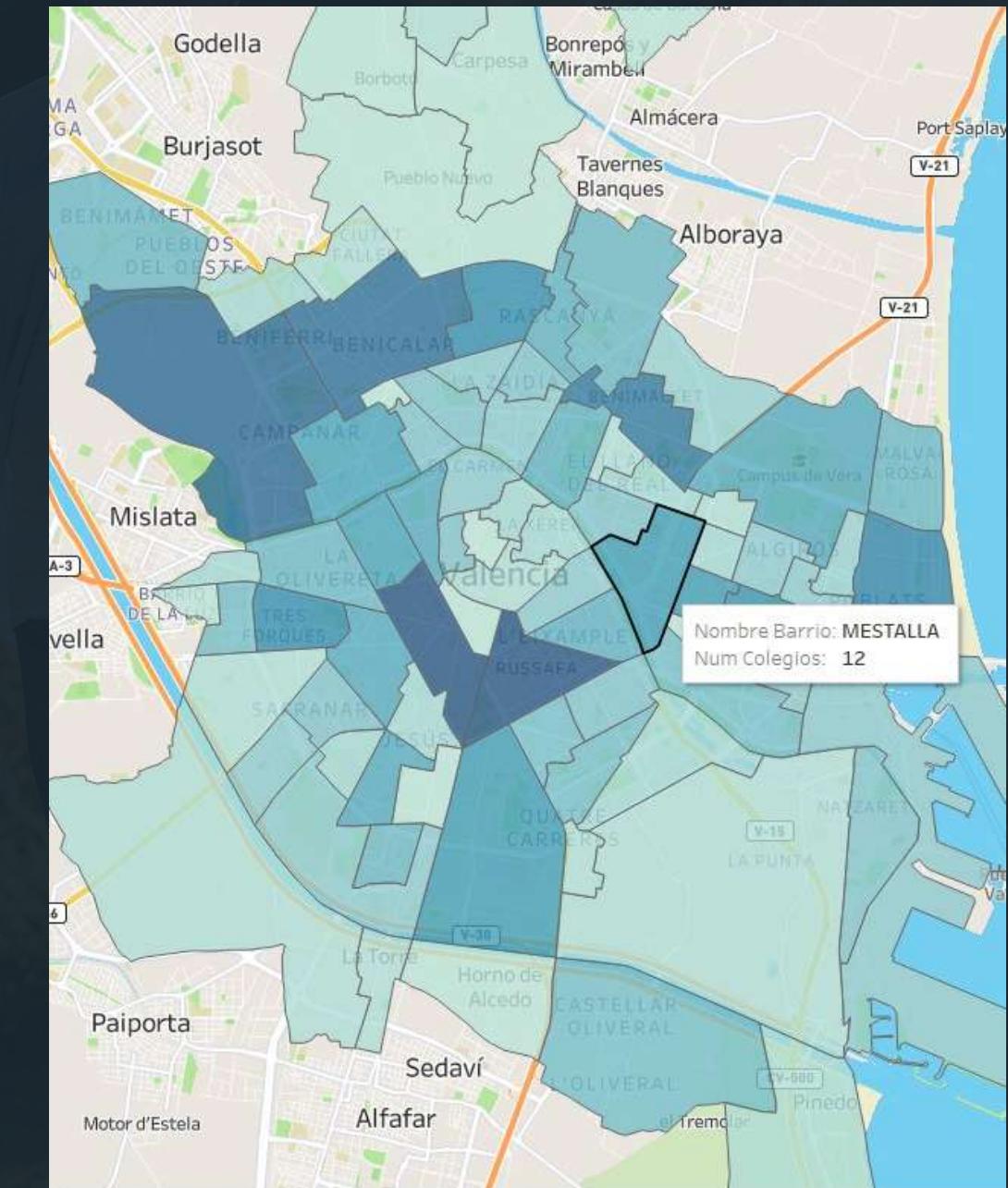
Zonas Verdes



Estaciones contaminación



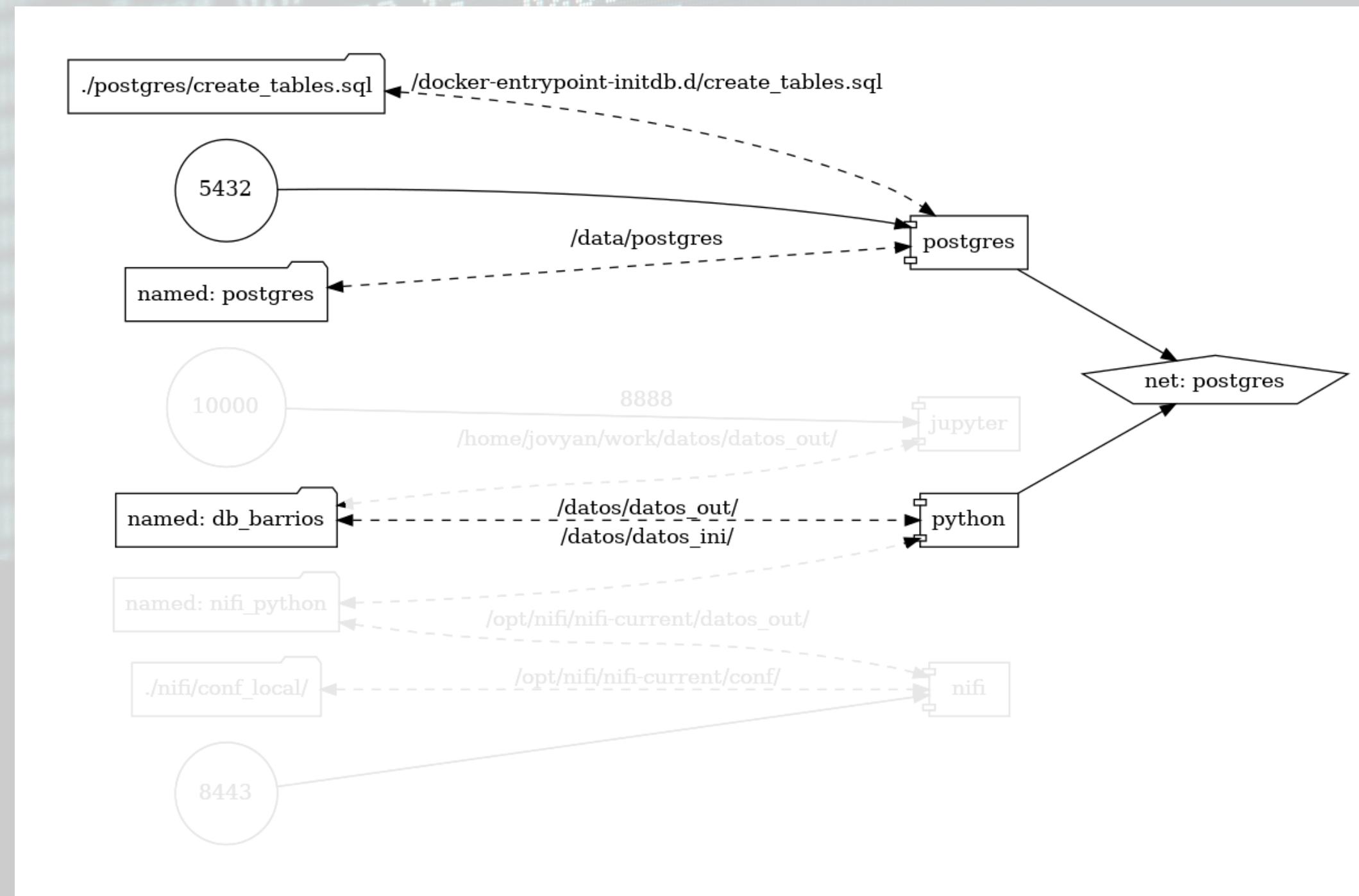
Centros Educativos



CONTENIDOS

- 
- 01 Introducción
 - 02 ETL calidad vida Valencia
 - 03 ETL clientes potenciales
 - 04 BD y calidad del dato
 - 05 Mejora y Visualización
 - 06 Conclusiones

¿Dónde estamos?



EXTRACCIÓN DE DATOS



Forma de obtención:

- **Encuesta sobre indicadores de calidad de vida**
 - **Transporte público**
 - **Centros educativos**
 - **Centros sanitarios**
 - **Zonas Verdes**
 - **Contaminación atmosférica**
 - **Contaminación acústica**
 - **Limpieza**
 - **Puntos de carga eléctricas**

API

```
def get_gform_clients(api_form,list_caract,nprefs):  
  
    import pandas as pd  
    from pydrive.auth import GoogleAuth  
    from pydrive.drive import GoogleDrive  
  
    gauth = GoogleAuth()  
    gauth.LoadCredentialsFile("./modulos/mycreds.txt")  
    if gauth.credentials is None:  
        gauth.LocalWebserverAuth()  
    elif gauth.access_token_expired:  
        gauth.Refresh()  
    else:  
        gauth.Authorize()  
    gauth.SaveCredentialsFile("./modulos/mycreds.txt")  
    drive = GoogleDrive(gauth)
```

Forma de obtención:

- **Conexión a la API de Google:**
 - **Autentificación de credenciales**
 - **Almacenamiento de credenciales**
 - **Descarga de datos desde la API de Google Sheets**

API

Conversión de archivos:

- **Creación de archivos en Google Drive a partir del ID de Google Sheets**
- **Cambio de formato de un archivo .xml a un spreadsheetml**
- **Por último, pandas lee el archivo spreadsheetml como un archivo .xls y crea un DataFrame**

```
file_obj = drive.CreateFile({'id': '${API_id}'})
file_obj.GetContentFile(api_form,
|   |   mimetype='application/vnd.openxmlformats-officedocument.spreadsheetml.sheet')

df = pd.read_excel(api_form)
```

CLIENTES POTENCIALES

Clientes ID	Salud	Limpieza	Colegios	Transp. Público	Estaciones de carga coche	PM25	Ruido	Zonas Verdes
1	int64	int64	int64	int64	int64	float64	float64	float64
2	int64	int64	int64	int64	int64	float64	float64	float64
3	int64	int64	int64	int64	int64	float64	float64	float64
...	int64	int64	int64	int64	int64	float64	float64	float64
n	int64	int64	int64	int64	int64	float64	float64	float64

PREFERENCIAS POR CLIENTE

Clientes ID	Preferencia 1	Preferencia 2	Preferencia 3
1	str	str	str
2	str	str	str
3	str	str	str
...	str	str	str
n	str	str	str

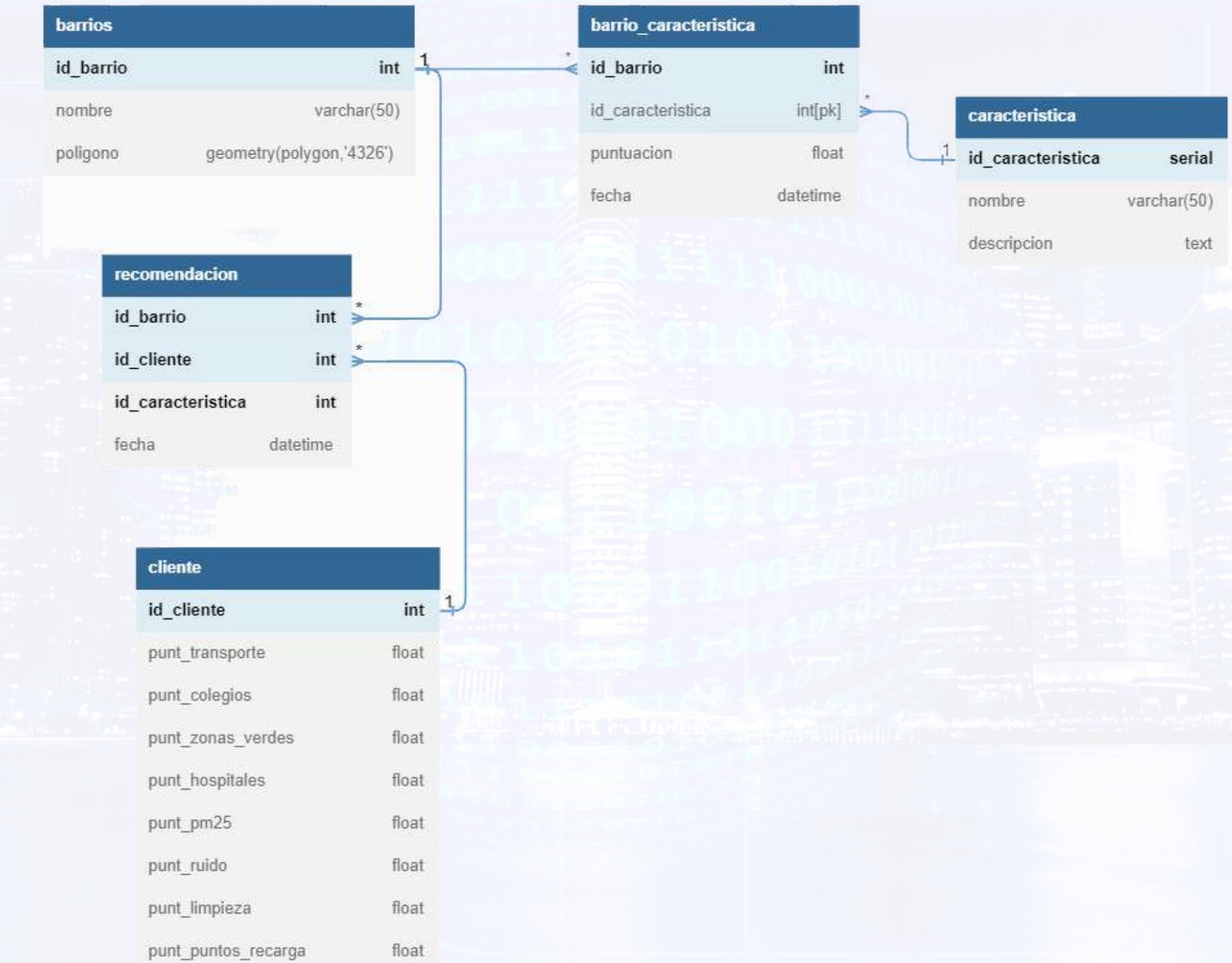
RECOMENDACIÓN CLIENTES

Ciudad ID	Barrio ID	Característica ID	Fecha
1	1º Barrio ID	ID Característica 1	D/M/A
1	2º Barrio ID	ID Característica 1	D/M/A
1	3º Barrio ID	ID Característica 1	D/M/A
1	4º Barrio ID	ID Característica 2	D/M/A
1	5º Barrio ID	ID Característica 2	D/M/A
...	nº Barrio ID	...	D/M/A
1	9º Barrio ID	ID Característica 3	D/M/A

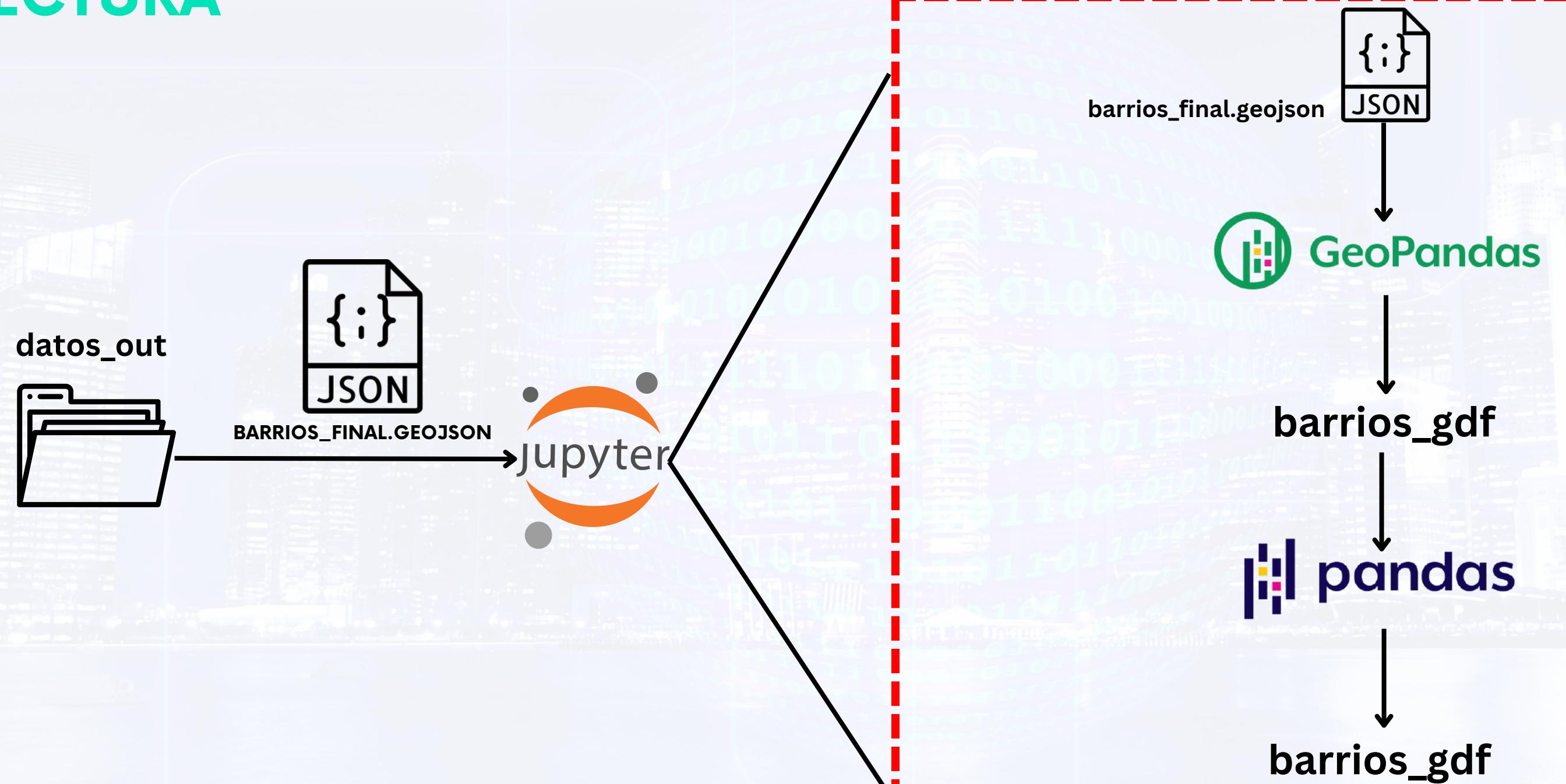
CONTENIDOS

- 
- 01 Introducción
 - 02 ETL calidad vida Valencia
 - 03 ETL clientes potenciales
 - 04 BD y calidad del dato
 - 05 Mejora y Visualización
 - 06 Conclusiones

BASE DE DATOS



LECTURA

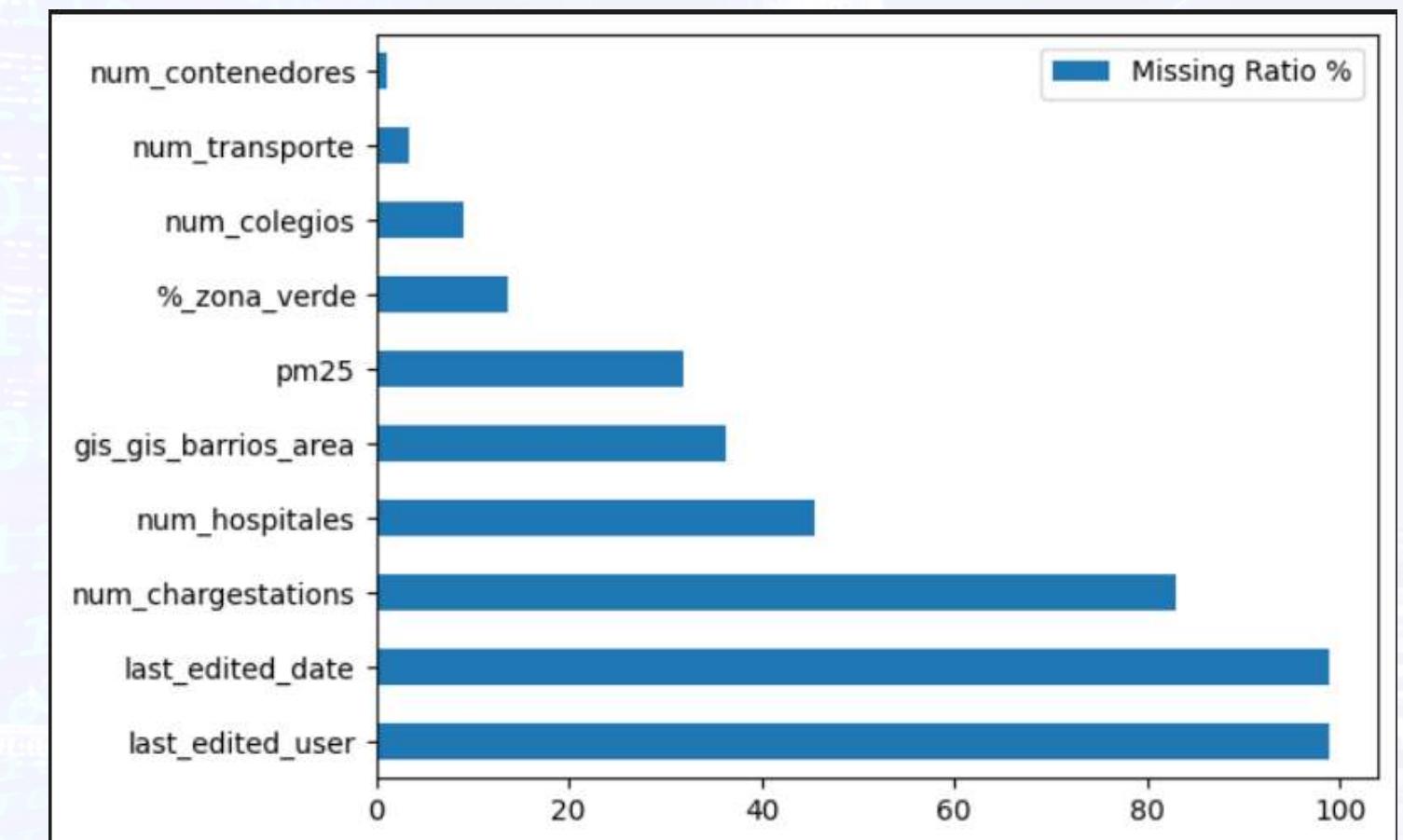


EXPLORACIÓN

Columnas

```
Index(['geometry', 'coddistrit', 'gis_gis_barrios_area', 'object_id_barrio',
       'linkid', 'codbarrio', 'coddistbar', 'geo_point_2d', 'nombre_barrio',
       'last_edited_user', 'last_edited_date', '%_zona_verde',
       'id_caract_%_zona_verde', 'nivel_acustico', 'id_caract_nivel_acustico',
       'num_hospitales', 'id_caract_num_hospitales', 'num_colegios',
       'id_caract_num_colegios', 'num_chargestations',
       'id_caract_num_chargestations', 'pm25', 'id_caract_pm25',
       'num_contenedores', 'id_caract_num_contenedores', 'num_transporte',
       'id_caract_num_transporte'],
```

Nulos



ELIMINACIÓN DE FILAS Y COLUMNAS

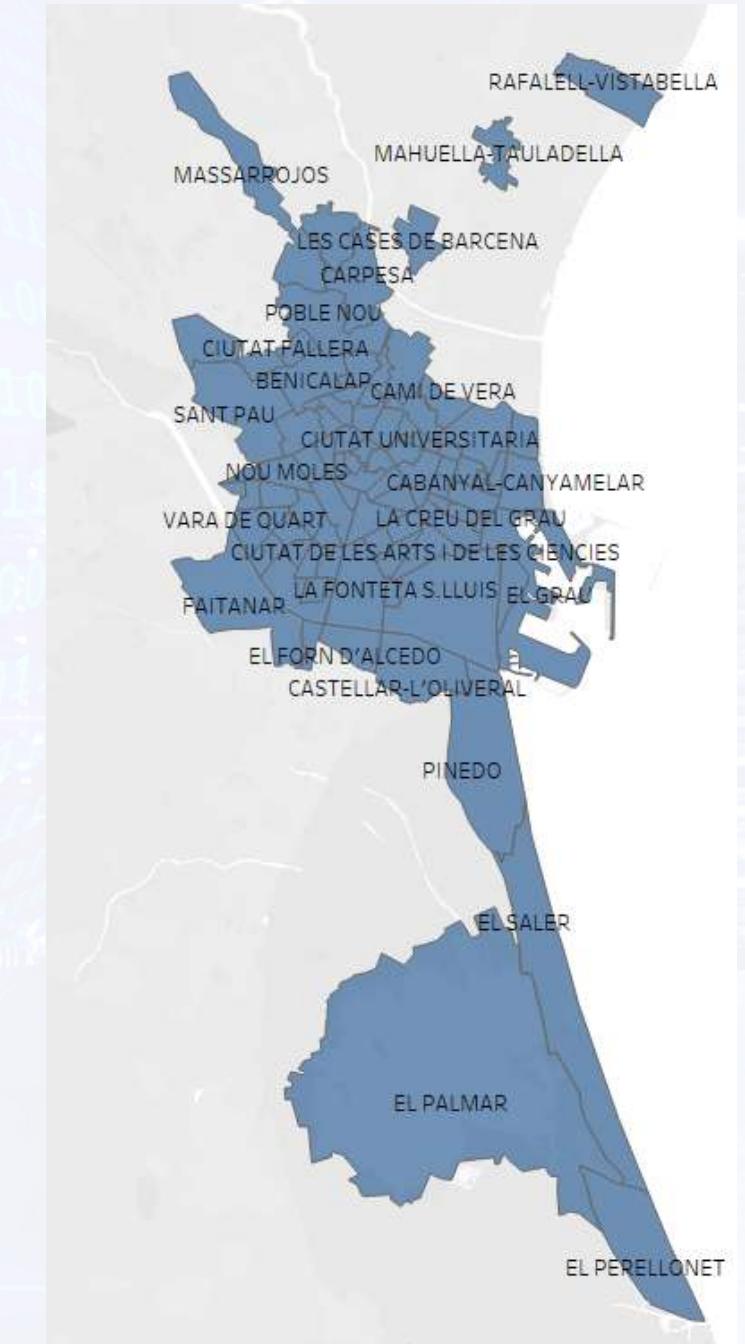
Columnas

```
Index(['geometry', 'coddistrit', 'gis_gis_barrios_area', 'object_id_barrio',
       'linkid', 'codbarrio', 'coddistbar', 'geo_point_2d', 'nombre_barrio',
       'last_edited_user', 'last_edited_date', '%_zona_verde',
       'id_caract_%_zona_verde', 'nivel_acustico', 'id_caract_nivel_acustico',
       'num_hospitales', 'id_caract_num_hospitales', 'num_colegios',
       'id_caract_num_colegios', 'num_chargestations',
       'id_caract_num_chargestations', 'pm25', 'id_caract_pm25',
       'num_contenedores', 'id_caract_num_contenedores', 'num_transporte',
       'id_caract_num_transporte'],
```

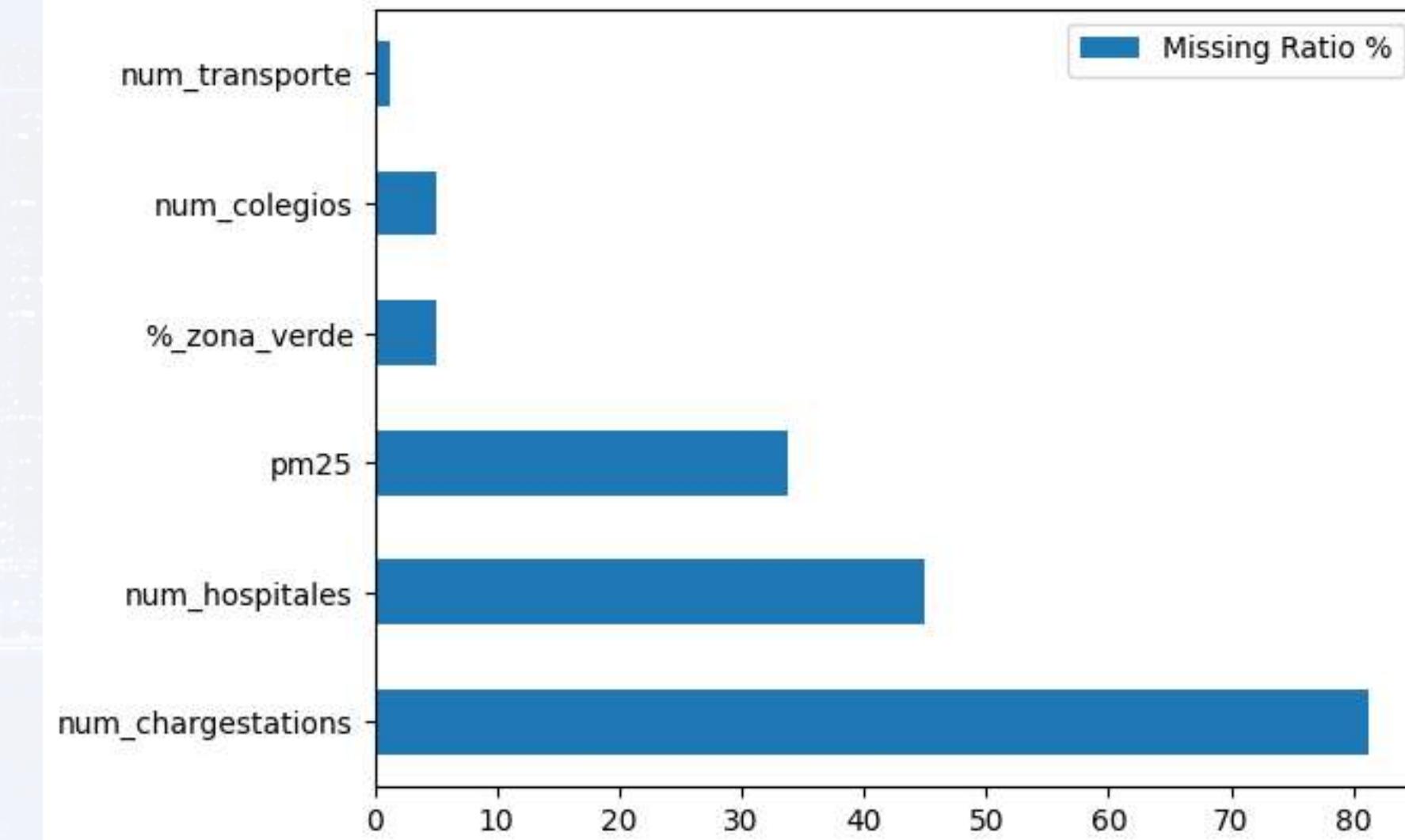


```
Index(['geometry', 'object_id_barrio', 'geo_point_2d', 'nombre_barrio',
       '%_zona_verde', 'id_caract_%_zona_verde', 'nivel_acustico',
       'id_caract_nivel_acustico', 'num_hospitales',
       'id_caract_num_hospitales', 'num_colegios', 'id_caract_num_colegios',
       'num_chargestations', 'id_caract_num_chargestations', 'pm25',
       'id_caract_pm25', 'num_contenedores', 'id_caract_num_contenedores',
       'num_transporte', 'id_caract_num_transporte'],
```

Filas



CAMBIO DE TIPOS DE DATOS Y TRATAMIENTO DE NULOS



CAMBIO DE TIPOS DE DATOS Y TRATAMIENTO DE NULOS

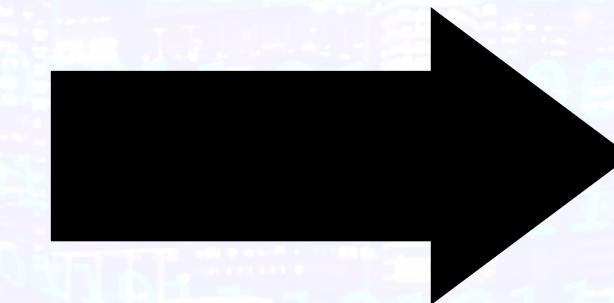
Rellenamos NaN con 0s y...

Fila	tipo de dato incial	tipo de dato futuro
Nº Hospitales	float64	int64
Nº Papeleras	float64	int64
Nº Colegios	float64	int64
Nº estaciones de transporte público	float64	int64
Nº Estaciones de carga coche	float64	bool

CAMBIO DE TIPOS DE DATOS Y TRATAMIENTO DE NULOS

Situación incial

	nombre_barrio	num_chargestations
0	NATZARET	NaN
1	LA CREU COBERTA	2.0
2	CIUTAT JARDI	NaN
4	LA CARRASCA	NaN
7	SANT LLORENS	NaN
...
82	BENICALAP	NaN
83	LA PETXINA	NaN
84	SANT FRANCESC	2.0
86	ELS ORRIOLS	NaN
87	NA ROVELLA	NaN

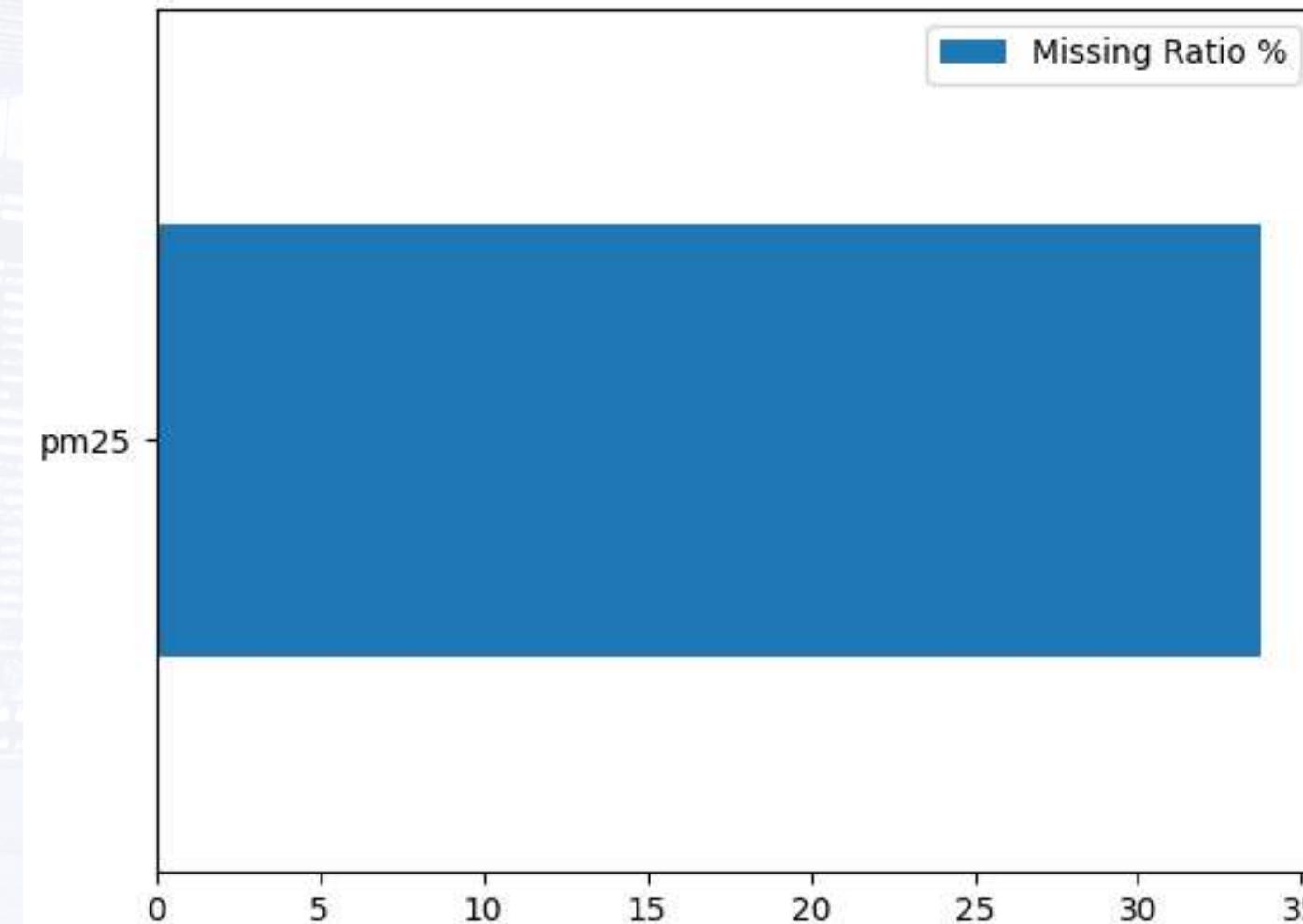


Situación final

	nombre_barrio	chargestations
0	NATZARET	False
1	LA CREU COBERTA	True
2	CIUTAT JARDI	False
4	LA CARRASCA	False
7	SANT LLORENS	False
...
82	BENICALAP	False
83	LA PETXINA	False
84	SANT FRANCESC	True
86	ELS ORRIOLS	False
87	NA ROVELLA	False

CAMBIO DE TIPOS DE DATOS Y TRATAMIENTO DE NULOS

Contaminación (pm25)



¿Cómo eliminamos estos nulos?

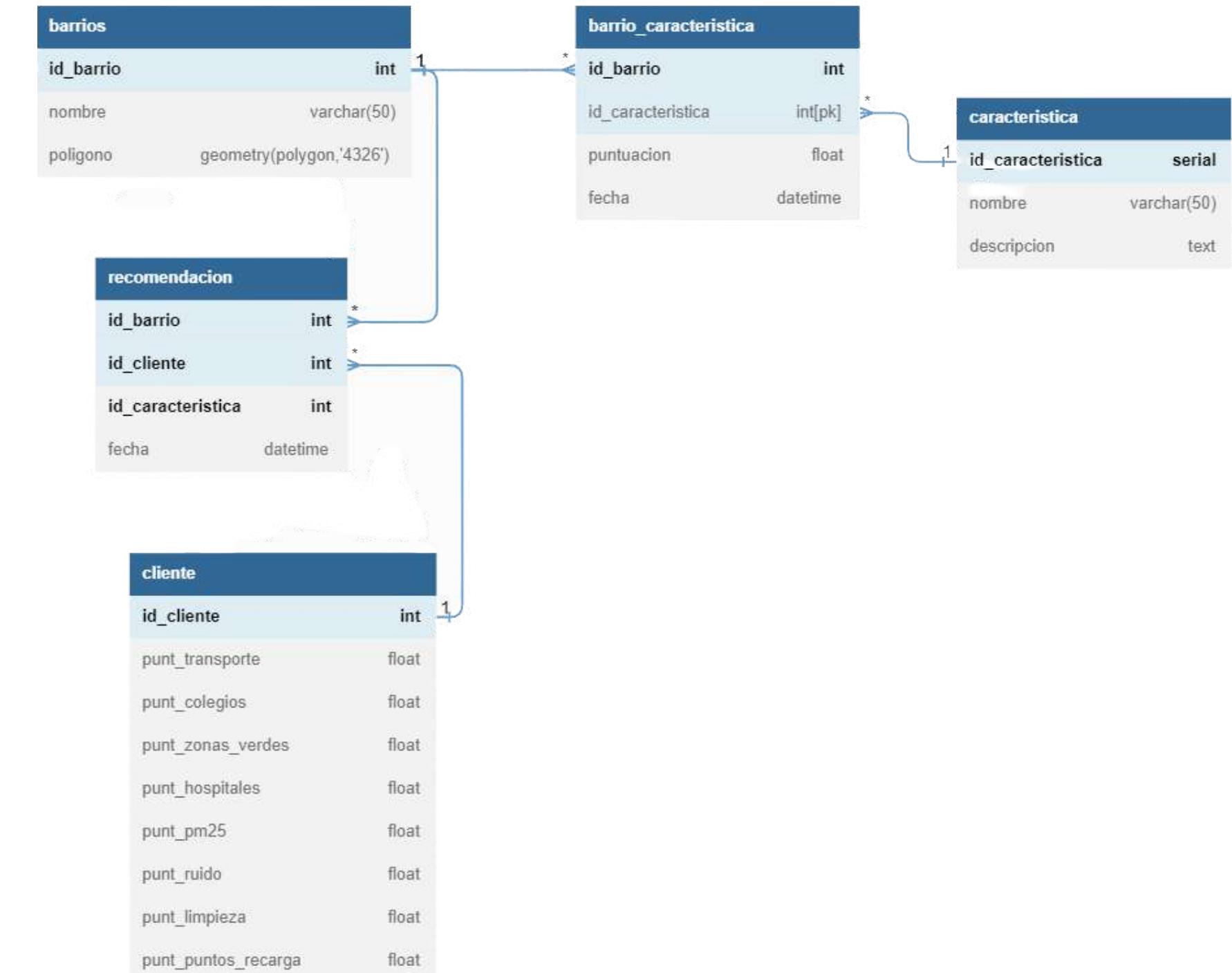
- Asignamos los barrios a una zona de Valencia (norte, sur, este o centro)
- Buscamos un índice de pm25 conocido para cada una de las zonas
- Extendemos ese índice de contaminación al resto de barrios que comparten zona

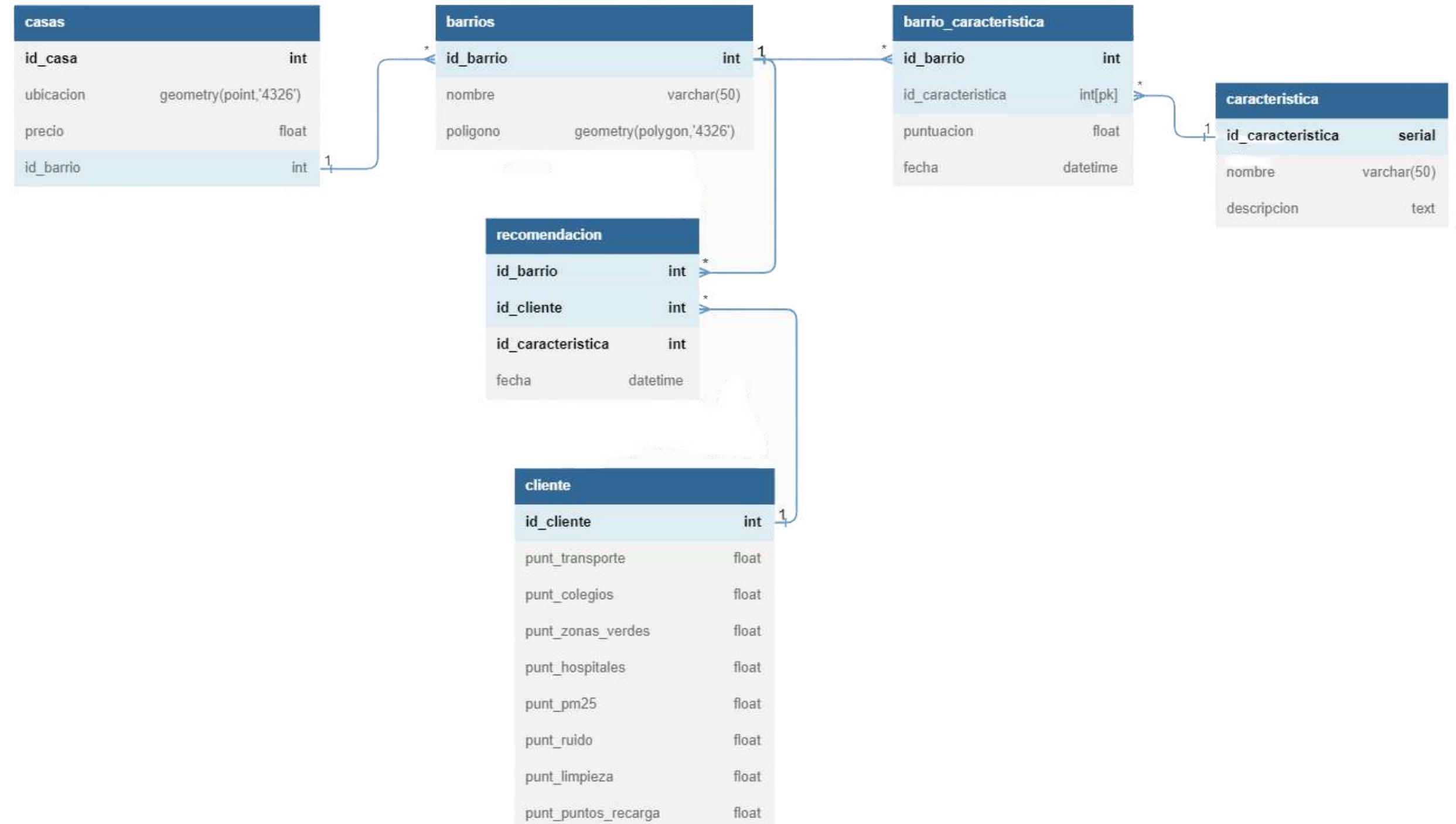
CONTENIDOS

- 
- 01 Introducción
 - 02 ETL calidad vida Valencia
 - 03 ETL clientes potenciales
 - 04 BD y calidad del dato
 - 05 Mejora y Visualización
 - 06 Conclusiones



UN PASO MÁS ALLÁ





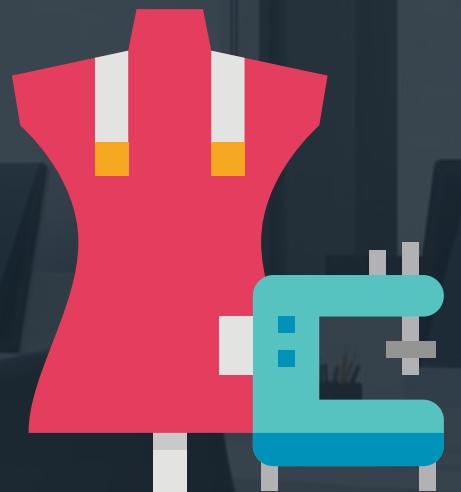
PRINCIPIOS



Relación calidad - precio



Visión completa



Personalización



Precisión



idealista
USUARIO

idealista
EMPRESA



LA HORA DE LA VERDAD

A photograph of several students in a classroom setting, focused on their work. In the foreground, a student's hands are visible on a laptop keyboard. Behind them, another student holds a tablet, and a third student is writing in a notebook with a pen. The scene conveys a sense of active learning and digital education.

Idealista
GRACIAS