DATA ANALYSIS OF THE SINGLE REGISTRY OF PUBLIC SERVICE PROVIDERS IN COLOMBIA

ESTEBAN COBO GOMEZ

JOSE DAVID MESA RAMIREZ

CARLOS ANDRES OROZCO

SARA LUCIA ROJAS MEJIA

BREYNER POSSO BAUTISTA

UNIVERSIDAD AUTÓNOMA DE OCCIDENTE

FACULTAD DE INGENIERÍA

INGENIERÍA DE DATOS E INTELIGENCIA ARTIFICIAL

ETL

Summary

This project uses an ETL (Extract, Transform, Load) process to work with data from the Unified Registry of Public Service Providers (RUPS) in Colombia. The objective is to identify how water supply, sewerage, and waste collection services are distributed and in what conditions they are provided across the country's municipalities.

Starting from a file with over 13,000 original rows, quality filters were applied, and the data was cleaned and organized, resulting in 9,286 valid records representing service providers by location. This enabled the generation of indicators (KPIs) and visualizations that help understand where there is full coverage, where environmental issues may exist (such as water service without sewerage), and which areas require more attention to meet SDG 6.

The result is a SQLite database (rups.db), several CSV reports, visualizations, and a structure ready for more advanced analyses such as mapping or regional comparisons.

Introduction and Context — SDG 6

In Colombia, access to clean water, basic sanitation (wastewater management), and proper waste collection remains unequal in many areas, particularly in rural or under-resourced municipalities. These three services are directly tied to public health, environmental protection, and social well-being.

This project seeks to analyze these services using real, open government data to identify which areas of the country have all three services, which have only one or two, and where there may be major public health or environmental risks affecting the quality of life in communities.

The analysis is aligned with Sustainable Development Goal 6 (SDG 6), proposed by the United Nations in its 2030 Agenda for Sustainable Development:

> "Ensure availability and sustainable management of water and sanitation for all."
> (UN, 2015)

From this perspective:

- **Water supply** refers to clean water delivered to households, essential for consumption, hygiene, and food safety.

- **Sewerage** is the system that collects and sometimes treats wastewater to prevent contamination of rivers, streams, or aquifers.

- **Waste collection**, while not directly involving water, significantly impacts it. It is included in this analysis because poor solid waste management (e.g., trash accumulation, illegal dumping, or lack of technical disposal) can produce leachate, blockages, or build-ups that ultimately contaminate nearby water bodies, harm the soil, and spread disease.

For these reasons, the project includes not just water supply and sewerage, but also waste collection services, recognizing that all three are part of a holistic view of water management in Colombia. This decision strengthens the project's contribution to SDG 6 by considering all ways in which water resources may be affected by inadequate public services.

General Objective

To design and apply an ETL process to the RUPS data in order to generate indicators, reports, and visualizations that reveal how water supply, sewerage, and waste collection services are distributed across Colombia, contributing to the monitoring of SDG 6.

Dataset Justification: What Does Each Row Represent?

The data comes from the public file:
"Registro_Único_de_Prestadores_de_Servicios_Públicos–RUPS.csv".

Each row represents a public service provider operating in a specific location (municipality and department). Since a single provider can operate in multiple zones, it's important to preserve the granularity by municipality to know exactly where each service is offered.

After applying quality filters and removing duplicate records (based on provider name and location), 6,848 rows were retained for the final analysis.

Key columns in the file include:

- **NIT**: Legal identification number of the provider.
- **NOMBRE** and **RAZON_SOCIAL**: Commercial and legal names of the provider.
- **SERVICIO** and **SERVICIO_DETALLE**: Indicate whether the provider offers water, sewerage, waste collection, or combinations.
- **DEPARTAMENTO_PRESTACION** and **MUNICIPIO_PRESTACION**: Exact location of service.
- **ESTADO**: Indicates whether the provider is active, canceled, or in another status.

- **FECHA_REGISTRO**: Date the provider was registered.

Other fields specify whether the provider is public or private, their target audience (residential, commercial), legacy system origin, and regulatory supervision.

During transformation, new columns were added to facilitate analysis:

- **has_acueducto**, **has_alcantarillado**, and **has_aseo**: Marked with 1 if the service is provided, 0 otherwise.
- **clasificacion_servicios**: Groups providers based on which services they offer (e.g., “AAA” if they provide all three).

Missing values were not deleted, as they help highlight areas where data is incomplete or poorly reported. The column **has_alcantarillado** was taken as the main proxy for sanitation coverage, as it best reflects real-world conditions.

Data Transformation and Quality Criteria

Several steps were taken during the ETL process to ensure data usability and reliability:

- All records were checked for valid department and municipality entries.
- Service columns were validated to match declared details.
- Duplicate records were removed to avoid double-counting.
- Regular expressions (regex) were used to detect services despite typos or inconsistent naming.
- Municipal-level detail was preserved for accurate territorial planning.

All steps were documented in a quality report to ensure transparency and traceability.

Why This Tech Stack?

The project was built using **Python 3.11**, a popular language for data science due to its ease of use, large community, and mature ecosystem.

- **pandas** was used for efficient data reading, transformation, and analysis.
- **SQLite** was chosen for storing the final data due to its lightweight, serverless structure. It uses a single .db file (rups.db), ideal for sharing and scaling later to PostgreSQL if needed.
- Visualizations were created using **matplotlib** and **seaborn**, which allow for clear and customizable charts.
- **pathlib** was used for safe, cross-platform path management.
- The **csv** module was used to export results in open, accessible formats.

In summary, the tools were chosen for being free, open-source, and highly replicable—allowing the project to run on any computer without licenses or expensive infrastructure.

Architecture Design

The project follows a modular ETL architecture (Extract, Transform, Load) that organizes the workflow in a clean and efficient way:

1. **Extraction**: Reads a CSV file with 13,000+ records.
2. **Transformation**: Applies quality filters, generates new features, and cleans data.
3. **Load**: Loads transformed data into a relational SQLite database (rups.db).
4. **Analysis**: KPIs and visualizations are generated using Python notebooks.

This architecture ensures each stage can be independently maintained or improved.

Data Model

Fact Table: fact_prestacion

The core of the star schema is the fact table fact_prestacion, which stores service provision records. Each row represents a unique combination of provider, location, services, and status.

- fact_id: Unique fact identifier
- prestador_id: FK to the provider dimension
- ubicacion_id: FK to the location dimension
- servicio_id: FK to the service dimension
- estado_id: FK to the provider status dimension

Dimensions

- dim_prestador: Provider information (name, NIT, type, classification)
- dim_ubicacion: Location (department, municipality)
- dim_servicio: Services (name, binary indicators for water, sewerage, cleaning)
- dim_estado: Operational/legal status (active, canceled, type of registration)

Key Performance Indicators (KPIs) — Coverage and Sanitation

Several KPIs were generated from the transformed RUPS data:

1. **General Summary**: 13,000+ records, covering all 32 departments and 1,000+ municipalities. Around 14% of providers offer AAA services.

2. **National Coverage**: Heatmaps by department and municipality reveal high- and low-density zones and gaps in dual service (water + sewerage).

3. **Record Density**: Calculates number of providers per municipality, revealing areas of high operational concentration vs. underserved regions.

4. **Water vs. Sewerage Analysis**: Categorizes municipalities into: both services, only water, only sewerage, or neither—exposing health risks when sewerage is missing.

5. **AA Coverage Rate**: Shows what % of municipalities in each department have both water and sewerage services.

6. **Georeferenced KPI**: Heatmap generated by associating latitude/longitude with each municipality + department, helping visualize service density and coverage gaps.

7. **Service Totals per Department**: Aggregates all services to highlight regions with high vs. low public service provision.

Together, these KPIs offer a comprehensive view of Colombia's water, sanitation, and cleaning services, exposing territorial inequalities and guiding public policy decisions in line with SDG 6.

Exploratory Data Analysis (EDA)

The exploratory analysis aimed to understand the distribution and classification of public services reported in the RUPS dataset, and to diagnose why the AAA coverage indicator (Water + Sewerage + Cleaning) initially showed 0%, despite clear evidence of these services in the data.

After loading and exploring the rups.db database, it was confirmed that the transformed dataset contains 9,286 valid records, representing service providers by municipality and department. This level of detail preserves the geographic granularity required for assessing actual service coverage across Colombia.

From the SERVICIO column, three binary flags were generated: has_acueducto, has_alcantarillado, and has_aseo. These indicate whether each service is provided per row. A new column (service_class) was created to classify service combinations into categories like "Water only", "AA", "AAA", and others.

Negative correlations were found between services (e.g., -0.72 between water and cleaning), suggesting that many providers specialize in a single service, which reduces the frequency of comprehensive (AAA) service delivery.

This EDA validated the classification logic, corrected the AAA KPI, and clarified the structure of the dataset—providing a solid foundation for further analysis, including service coverage, classification, and geospatial insights.

Conclusions

This project implemented a full ETL workflow on public RUPS data to analyze the distribution of water, sewerage, and waste services in Colombia. By cleaning and structuring over 13,000 records, a relational database was built to support the generation of key indicators aligned with SDG 6.

The KPIs reveal significant territorial disparities—such as municipalities with water access but no sewerage, and a low proportion of providers with full (AAA) coverage. Geospatial analysis identified critical regions with low public service presence, posing challenges for equity and environmental sustainability.

Altogether, this work provides a solid foundation for future enhancements such as dashboards, interactive maps, or predictive models, and represents a technical and academic contribution to understanding essential public service coverage in Colombia.

References

United Nations. (2015). *Transforming our world: the 2030 Agenda for Sustainable Development*. https://sdgs.un.org/es/goals

GitHub**: https://github.com/EstebanC111s/ETL-Project-First-Delivery**

Slides Prezi:
https://prezi.com/view/HrJpRvnppcXUwNmW4a3S/?referral_token=ueXo55lnB3FN