PROYECTO
SUSTAINABLE
GROWTH MONITOR

Simulación Laboral

No Country

Entregable 1 y 2:

Definición de Métricas KPIs y Modelo de datos consolidado con fuentes simuladas.

2025

Team S11-25-Equipo66-BI:
Daniela Artica
Yira Marchitelli
Marianela Pi

Contenido

1. Introducción y Propuesta de Valor

El objetivo principal de este proyecto es diseñar un sistema de inteligencia de negocios que permita a las PyMEs monitorear su desempeño integral. Actualmente, la información financiera, operativa y de recursos humanos reside en silos desconectados.

Nuestra propuesta de valor consiste en **integrar estos indicadores** para demostrar que la sostenibilidad no es un centro de costos, sino una palanca de rentabilidad. El dashboard resultante permitirá responder preguntas clave como: ¿Cómo impacta la eficiencia energética en el margen neto? o ¿Qué relación existe entre la satisfacción del empleado y los costos de rotación?

Para hacer posible este análisis, se construyó un modelo de datos consolidado que integra y organiza todas las fuentes simuladas del proyecto, sirviendo como base estructural del dashboard y del análisis de correlaciones.

2. Definición de Métricas e Indicadores (KPIs)

A continuación, se detallan los indicadores seleccionados para cubrir los requerimientos financieros y ESG (Ambiental, Social y Gobernanza), junto con su lógica de cálculo y objetivos.

2.1 Métricas Financieras

Objetivo: Medir la rentabilidad y salud económica del negocio.

Indicador / KPI	Definición / Fórmula	Propósito de Negocio
Ingresos Totales	SUM(ingresos)	Medir el volumen de ventas facturado.
Costos Operativos	SUM(costos)	Controlar el gasto total de la operación.
Rentabilidad (Margen Neto)	SUM(ingresos) - SUM(costos)	Medir la ganancia real en dinero.
Margen Porcentual	([Margen Neto] / SUM(ingresos))	Evaluar la eficiencia y rentabilidad relativa del negocio.
Crecimiento de Ingresos	([Ingresos Actuales] - [Ingresos Periodo Anterior]) / [Ingresos Periodo Anterior]	Monitorear la expansión del negocio año contra año.

2.2 Métricas Ambientales (E)

Objetivo: Monitorear el impacto ecológico y la eficiencia de recursos.

Indicador / KPI	Definición / Fórmula	Propósito de Negocio	Meta Sugerida
Consumo Energético	AVERAGE(consumo_kwh)	Detectar picos de ineficiencia energética.	< 1800 kWh/día
Huella de Carbono	AVERAGE(huella_carbono_tCO2e)	Medir el impacto ambiental directo.	Reducción 10% anual
Consumo de Agua	AVERAGE(consumo_agua_litros)	Controlar el uso de recursos hídricos.	Reducción 5% anual
Tasa de Reciclaje	SUM(residuos_reciclados_kg) / SUM(residuos_totales_kg)	Verificar el compromiso con la economía circular.	> 40%

2.3 Métricas Sociales (S)

Objetivo: Evaluar el capital humano, la equidad y el clima laboral.

Indicador / KPI	Definición / Fórmula	Propósito de Negocio	Meta Sugerida
Tasa de Rotación	(SUM(empleados_baja) / AVERAGE(total_empleados))	Medir la estabilidad de la fuerza laboral y costos asociados.	< 15% anual
Equidad de Género	AVERAGE(mujeres_liderazgo)	Monitorear la diversidad en puestos de decisión.	> 10 puestos clave
Satisfacción Empleado	AVERAGE(satisfaccion_empleados)	Evaluar el clima laboral y su impacto en la productividad.	> 7.5 puntos

2.4 Métricas de Gobernanza (G)

Objetivo: Asegurar la transparencia, ética y cumplimiento normativo.

Indicador / KPI	Definición / Fórmula	Propósito de Negocio	Meta Sugerida
Capacitación Ética	AVERAGE(pct_capacitacion_etica)	Mitigar riesgos legales y de reputación.	> 95%
Auditorías Internas	AVERAGE(nro_auditorias_internas)	Medir el control interno promedio por trimestre.	>= 2 por trimestre
Canal de Denuncias	LASTNONBLANK(canal_denuncias_activo, 1)	Garantizar mecanismos de transparencia activos.	100% Operativo

3. Estrategia de Alertas y Objetivos

Para cumplir con el requerimiento funcional de "detectar desviaciones", el sistema no sólo medirá el valor actual, sino que lo comparará contra **Objetivos Anuales Definidos**.

- **Mecanismo:** Cada métrica listada anteriormente tendrá asociado un "Valor Meta" para el año en curso.
- **Visualización:** El dashboard resaltará automáticamente (en rojo/verde) cuando un KPI se desvíe de su meta establecida (ej. si la Tasa de Reciclaje cae por debajo del 40%).

4. Arquitectura del Modelo de Datos

El modelo de datos del proyecto Sustainable Growth Monitor se diseñó siguiendo un enfoque estrella (Star Schema), optimizado para análisis de Business Intelligence. La arquitectura está compuesta por:

4.1. Una dimensión temporal limpia (dim_tiempo)

Proporciona la granularidad diaria y los atributos necesarios para realizar:

- agregaciones por trimestre,
- análisis por año, mes o día,
- cálculo de correlaciones temporales.

La clave primaria **id_fecha** se utiliza como punto de unión entre todas las tablas de hechos. Esta dimensión **no permite valores nulos en ninguno de sus campos**, ya que la estructura temporal debe ser completamente consistente para evitar errores en correlaciones y cálculos derivados.

4.2. Cuatro tablas de hechos operativas

Cada área del negocio se modela de forma independiente según su naturaleza:

• fact_finanzas

Registra ingresos y costos diarios. Permite datos incompletos porque simula escenarios reales donde la información financiera puede contener errores o ausencias de valores.

Campos que permiten nulos:

- **ingresos**: se habilitó para representar registros contables faltantes o inconsistencias operativas.
- **costos**: *no* admite nulos para garantizar un mínimo de completitud en la información financiera.

• fact_ambiental

Contiene consumo energético, agua, residuos y huella de carbono. También admite nulos, reflejando la variabilidad y calidad irregular típica de mediciones ambientales.

Campos que permiten nulos:

- **consumo_agua_litros**: simula sensores fuera de línea o fallas en medición.
- **huella_carbono_tCO2e**: permite nulos para reflejar cálculos incompletos o no reportados.

Los demás campos son obligatorios para asegurar un mínimo de coherencia ambiental diaria.

• fact_rrhh

Integra indicadores sociales: total de empleados, entradas, salidas, liderazgo y satisfacción. Representa el componente “S” de ESG.

Campos que permiten nulos:

- **satisfaccion_empleados**: se permite para simular encuestas no respondidas o periodos sin medición.

Todos los demás campos son obligatorios para reflejar consistencia en las métricas de personal.

• fact_gobernanza_trimestral

Modelada con granularidad trimestral, dado que las prácticas de gobernanza no ocurren diariamente. Su unión a *dim_tiempo* se hace mediante una fecha representativa del trimestre (por ejemplo, el último día del trimestre).

Campos que permiten nulos:

- **pct_capacitacion_etica**: puede faltar en algunos trimestres para simular reportes incompletos.

Los demás campos son obligatorios porque su ausencia comprometería la interpretación de auditorías y del estado del canal de denuncias.

4.3. Una tabla de catálogo independiente (objetivos_esg)

Esta tabla no forma parte del modelo relacional analítico, sino que funciona como un repositorio descriptivo de KPIs:

- Nombre del indicador
- Unidad de medida
- Definición
- Valor objetivo
- Año objetivo
- Ejemplo de valor típico

No requiere relaciones físicas, ya que su uso es conceptual y su contenido permite que se consuma desde la capa BI para comparar valores reales vs metas.

Campos que permiten nulos:

- **descripcion_objetivo**, **unidad_medida** y **valor_ejemplo**, ya que algunos objetivos pueden no requerir detalles adicionales o formatos específicos.

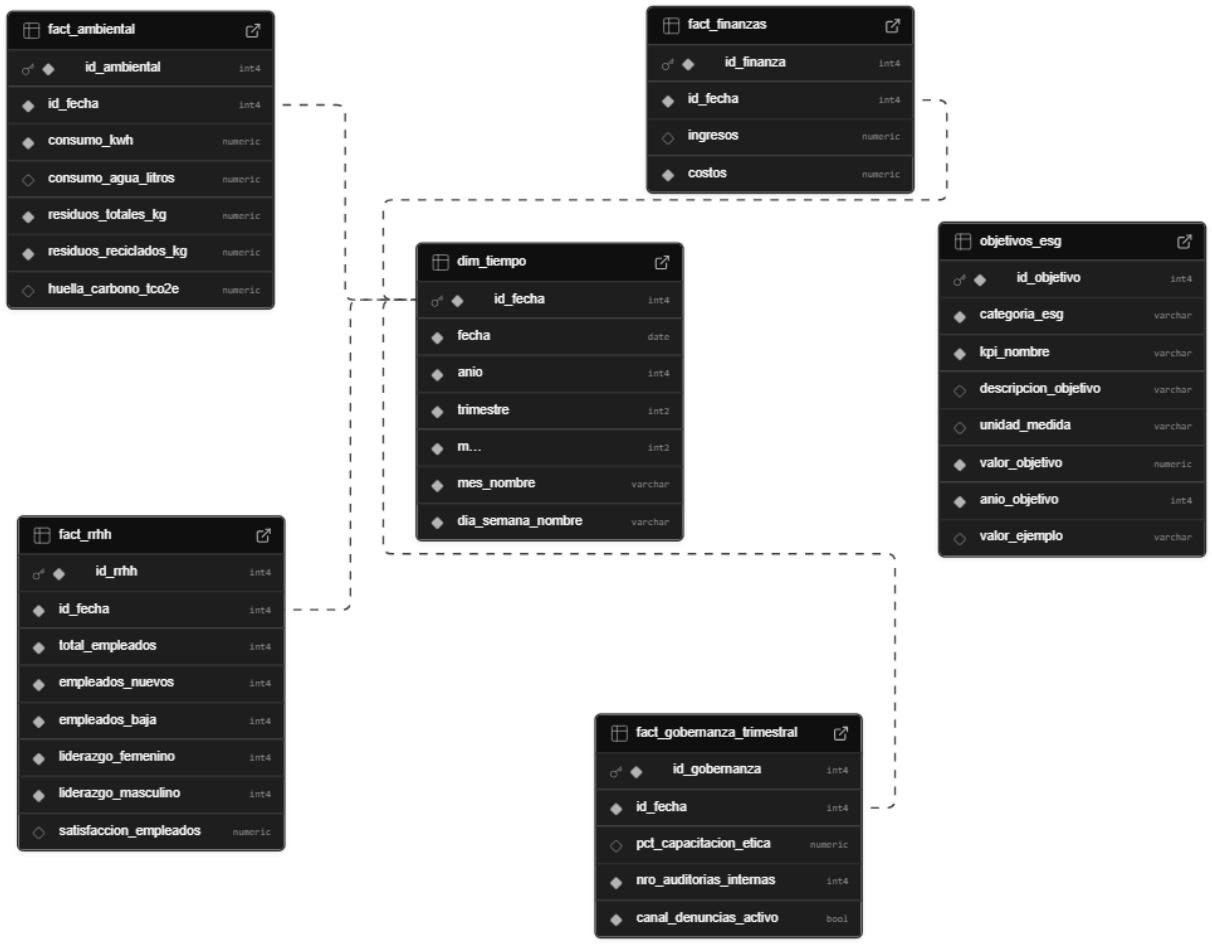
Existe como tabla independiente porque no participa en joins operativos, sino en la interpretación de KPIs.

4.4. Justificación de la arquitectura

Este diseño:

- Mantiene el modelo simple y entendible para analistas.
- Facilita cálculos temporales y correlaciones.
- Permite cargar datos imperfectos para simular ambientes reales.
- Reduce la rigidez del esquema permitiendo nulos en las tablas de hechos solo en campos donde los errores son realistas.
- Ofrece una tabla auxiliar (*objetivos_esg*) que sirve a la capa BI sin sobrecargar el esquema relacional.

4.5. Diagrama E-R:



5. Fuentes Simuladas y Estrategias de Generación

Este apartado describe detalladamente cómo fueron generadas las fuentes de datos que conforman el modelo. Incluye la lógica aplicada, el propósito de cada conjunto de datos y las técnicas utilizadas para simular escenarios reales con datos imperfectos.

5.1. Descripción General del Proceso

Los datos del proyecto fueron generados mediante un script en Python diseñado específicamente para:

- Crear fuentes simuladas con comportamientos realistas.
- Representar dinámicas financieras, ambientales, sociales y de gobernanza.
- Incluir **errores controlados** (nulos, duplicados y outliers) para evaluar limpieza, validaciones y análisis.
- Mantener **coherencia estructural** con la arquitectura del modelo, especialmente respecto a:

    - La **dimensión temporal limpia**.
    - La granularidad diaria y trimestral definida en las tablas de hechos.

El proceso produce **6 archivos CSV**, los cuales fueron posteriormente cargados en la base de datos.

5.2. Generación de la tabla dim_tiempo (Fuente limpia)

- Se generaron todas las fechas desde **01/01/2023 hasta 31/12/2025**.
- No contiene errores, nulos ni outliers.
- Incluye atributos derivados: trimestre, mes, nombres y día de la semana.
- Esta tabla es el eje temporal del modelo y la única diseñada para ser perfectamente limpia.

Motivo de limpieza absoluta:
Una dimensión temporal debe ser completa y estable, ya que todas las tablas fact dependen de una fecha válida.

5.3. Generación de la tabla fact_finanzas

Incluye métricas diarias de ingresos y costos.

Lógica de simulación:

- Ingresos: 10,000–50,000.
- Costos: 5,000–25,000.

Errores introducidos:

- **30 valores nulos en ingresos** → porque en esta tabla se permite nulos en este campo.
- **3 ingresos extremadamente altos (9,999,999)** → simulación de outliers.
- **10 registros duplicados** → simulan fallos típicos de exportación o ETL.

Granularidad:

- **Diaria**, como establece la arquitectura general del modelo.

5.4. Generación de la tabla fact_ambiental

Incluye consumo energético, agua, residuos y emisiones.

Lógica de simulación:

- Consumo kWh: 1,000–3,000.
- Consumo de agua: 500–1,500 L.
- Residuos totales: 50–200 kg.
- Reciclaje: 10%–70% del total.
- Huella de carbono derivada del consumo energético.

Errores introducidos:

- **25 valores nulos en consumo_agua_litros**.
- **10 valores nulos en huella_carbono_tCO2e**.

- No se generaron duplicados, simulando una medición más estable.

Justificación:
Representan sensores fuera de línea, transmisiones fallidas y reportes incompletos.

Granularidad:

- **Diaria**.

5.5. Generación de la tabla fact_rrhh

Incluye métricas sociales: empleados totales, rotación, liderazgo y satisfacción.

Lógica de simulación:

- Total empleados: 40–60.
- Entradas, bajas y liderazgo generados con rangos realistas.
- Satisfacción: 6.5–8.5.

Errores introducidos (alineados con el modelo del punto 4):

- **30 valores nulos en satisfacción_empleados** (este campo sí permite nulos).
- **15 registros duplicados** → simulan errores comunes en RRHH.

Granularidad:

- **Diaria**.

5.6. Generación de la tabla fact_gobernanza_trimestral

Esta tabla opera con **granularidad trimestral**, tal como especifica la arquitectura.

Lógica de simulación:

- Se utiliza **una fecha representativa por trimestre** (fin del trimestre).
- Métricas:
    - % capacitación ética (70%–95%)
    - Auditorías internas (1–2)
    - Canal de denuncias (activo/inactivo)

Errores introducidos:

- **2 valores nulos** en capacitación ética.

Justificación de granularidad:
Los indicadores de gobernanza no se reportan diariamente, sino en ciclos más amplios.

5.7. Tabla objetivos_esg

Es un **catálogo corporativo de KPIs**, no un hecho ni una dimensión operativa.

Estructura simulada:
Incluye objetivos 2023–2025 para categorías:

- Ambiental
- Social
- Gobernanza
- Financiero

Cada objetivo contiene:

- Unidad de medida
- Valor objetivo
- Descripción
- Valor ejemplo

5.8. Consistencia global del diseño

- Todas las tablas fact comparten **id_fecha** para permitir joins consistentes.
- Gobernanza trabaja con granularidad trimestral, como se definió en el modelo.
- Los errores simulados coinciden con las reglas y nulos permitidos del punto 4.
- Todas las tablas fueron generadas en **CSV** para facilitar la carga y validaciones posteriores.

6. Proceso de Carga y Validación de Datos en la Base

6.1. Carga de datos en Supabase

La carga de los datos simulados se realizó desde **WSL (Ubuntu en Windows)** para disponer de herramientas nativas de PostgreSQL y facilitar la interacción con Supabase.

Los archivos CSV se encontraban almacenados en el sistema Windows y fueron accedidos desde WSL mediante su ruta equivalente en /mnt/c/. Antes de iniciar la carga, se verificó la presencia de los seis archivos correspondientes a las tablas del modelo de datos:
dim_tiempo, **fact_finanzas**, **fact_ambiental**, **fact_rrhh**, **fact_gobernanza_trimestral** y **objetivos_esg**.

Para automatizar el proceso y asegurar uniformidad en la ingestión, se creó un script de importación que recorre cada archivo y lo carga en su tabla correspondiente dentro del esquema **sustainable_growth** en Supabase. Este script fue almacenado en la carpeta designada para procesos de carga dentro del directorio de trabajo del proyecto.

La importación se llevó a cabo utilizando el cliente psql desde WSL, conectándose al servidor PostgreSQL gestionado por Supabase y ejecutando la instrucción necesaria para copiar cada archivo hacia su tabla correspondiente.

Durante el proceso se validó que:

- Cada tabla recibiera la cantidad de registros prevista.

- No se presentaran errores relacionados con tipos de datos o formato.
- Los nulos y duplicados intencionales definidos en la generación de los datos fueran aceptados correctamente según los permisos del esquema.
- La tabla **dim_tiempo**, al ser la única fuente completamente limpia, se cargara sin inconsistencias.

Al finalizar, todas las tablas fueron cargadas exitosamente, con los volúmenes esperados y sin errores críticos, dejando la base lista para las validaciones posteriores descritas en los siguientes apartados.

6.2. Validación técnica posterior a la carga

Se verificaron las siguientes condiciones estructurales:

- **Integridad referencial:**
  Todas las tablas fact enlazan correctamente a dim_tiempo mediante id_fecha sin producir errores de clave foránea.
- **Índices funcionales:**
  Se validó el correcto funcionamiento de los índices creados en cada tabla fact para consultas por fecha.
- **Consistencia de granularidad:**
  - Tablas financieras, ambientales y de RRHH: registros diarios.
  - Gobernanza: registros trimestrales con una fecha representativa válida en dim_tiempo.

6.3. Manejo de datos imperfectos

Los valores nulos, duplicados y outliers se **mantuvieron tal como fueron generados**, ya que forman parte del diseño de simulación del proyecto.

Desde la perspectiva del DBA:

- No se corrigieron datos en la base.
- La estructura permite nulos y duplicados sin comprometer la integridad.

7. Limpieza de Datos

A continuación, se detallarán las acciones y el razonamiento aplicado a cada una de las tablas fuente para garantizar la calidad, consistencia y fiabilidad de los datos utilizados en la construcción del dashboard de sostenibilidad. Cabe señalar que, inicialmente, se utilizó la data extraída mediante el script de conexión a la base de datos Supabase. Sin embargo, durante la fase de limpieza, se identificó que la tabla de dimensión de tiempo original presentaba un alcance temporal incompleto, lo cual generaría valores nulos críticos al intentar consolidar los datos de las tablas de hechos para los registros posteriores a esa fecha.

Por ello, se reemplazaron las cinco tablas de hechos y la tabla dim_tiempo con los archivos CSV proporcionados directamente por la Administradora de la Base de Datos (DBA). Estos archivos fueron pre-validados para asegurar que la tabla dim_tiempo fuera completa y que la granularidad de las tablas de hechos fuera consistente con los requisitos del proyecto.

Los pasos de limpieza y tratamiento de *outliers* detallados en este documento se aplican exclusivamente a los archivos CSV proporcionados por la DBA a partir de esta fase.

7.1. Limpieza de la tabla dim_tiempo

- **Estructura y consistencia:** La tabla se encontró consistente y bien estructurada, con 1096 registros que abarcaban el periodo de tiempo entre 2023-01-01 y 2025-12-31.
- **Tratamiento de tipos de datos:** Se convirtió la columna fecha de tipo object a tipo datetime para permitir la verificación de la secuencia temporal (detección de días faltantes) y facilitar la agregación de datos a niveles mensuales y trimestrales en pasos posteriores.
- **Verificación de secuencia:** Se verificó la secuencia de días para asegurar que la serie de tiempo estuviera completa y se pudiera alinear correctamente con las tablas de hechos.

7.2. Limpieza de la tabla finanzas

- **Verificación de duplicados:** Se detectaron 10 registros duplicados en la columna id_fecha.

  Decisión: Estos registros se eliminaron debido a que su presencia duplicaría artificialmente las métricas de ingresos y costos, lo cual sesgaría el margen neto. Se optó por mantener la primera instancia de la fecha para preservar el registro válido original.

- **Tratamiento de tipos de datos:** Se mantuvo la columna id_fecha como int64 para la relación con dim_tiempo. Se creó una columna auxiliar (fecha) convertida a datetime para los procesos de limpieza.
- **Tratamiento de Valores Nulos:** Se detectaron 31 valores nulos en la columna ingresos.

  Decisión: Se imputaron los valores nulos con cero (0). En el contexto de datos diarios, se asumió que un valor nulo en ingresos representa un día en el que no se registró ninguna venta o ingreso, lo cual es un valor de hecho significativo.

- **Tratamiento de Valores Atípicos (Outliers):** Se detectaron tres valores atípicos extremos en la columna ingresos mediante el método de rango intercuartílico (IQR) y visualización de diagramas de caja.

  Decisión: Se eliminaron los tres registros con valor 9,999,999.0 en la columna ingresos, dado que sesgarían drásticamente el cálculo de los KPIs y representaban menos del 0.5% de todo el dataset.

7.3 Limpieza de la tabla ambiental

- **Verificación de duplicados:** Se confirmó que no había filas duplicadas, manteniendo un total de 1095 registros después de la limpieza.
- **Tratamiento de valores nulos:** Se detectaron 25 registros nulos en la columna consumo_agua_litros y 10 registros nulos en la columna huella_carbono_tco2e.

Decisión: Se aplicó interpolación lineal basada en la fecha para imputar los valores faltantes, dado que los valores nulos estaban distribuidos de forma aislada en una serie temporal diaria. La interpolación permite rellenar los huecos manteniendo la tendencia natural del consumo (energía y agua) y la huella de carbono, evitando la pérdida innecesaria de datos.

- **Tratamiento de Valores Atípicos (Outliers):** Se detectó un solo valor atípico en la columna residuos_reciclados_kg mediante el método de rango intercuartílico (IQR) y visualización de diagramas de caja.

  Decisión: Se eliminó el registro que contenía el valor atípico, dado que este registro representa el 0.09% del conjunto de datos.

7.4. Limpieza de la tabla recursos humanos

- **Verificación de duplicados:** Se encontraron 15 registros duplicados en la columna id_fecha.

  Decisión: Estos registros se eliminaron debido a que su presencia podría sesgar los KPIs de esa área. Se optó por mantener la primera instancia de la fecha para preservar el registro válido original.

- **Tratamiento de Valores Nulos:** Se detectaron 31 valores nulos en la columna satisfaccion_empleados.

  Decisión: Se aplicó interpolación temporal basada en la fecha para imputar los valores faltantes, dado que la satisfacción de los empleados se trata de un indicador continuo y estable y los valores nulos estaban distribuidos de forma aislada en una serie temporal diaria.

- **Tratamiento de Valores Atípicos (Outliers):** No se detectaron valores atípicos en ninguna de las columnas del dataset al implementar el método del rango intercuartílico (IQR) y elaboración de diagramas de caja.

7.5. Limpieza de la tabla gobernanza

- **Verificación de duplicados:** Se confirmó que no había filas duplicadas, manteniendo un total de 12 registros después de la limpieza.
- **Granularidad:** La tabla solo contiene 12 registros (12 trimestres). Esto hizo que el tratamiento de nulos fuera extremadamente sensible.
- **Tratamiento de Valores Nulos:** Se detectaron 2 valores nulos en la columna pct_capacitacion_etica

  Decisión: Dada la baja cantidad de registros (12), eliminar 2 filas representaría una pérdida del 16% de los datos de la serie. Por ello, se aplicó interpolación lineal entre los valores conocidos más cercanos, evitando la interpolación temporal que asumiría continuidad en períodos no reportados.

- **Tratamiento de Valores Atípicos (Outliers):** Se detectó un valor atípico en la columna pct_capacitacion_etica mediante el método de rango intercuartílico (IQR) y visualización de diagramas de caja.

  Decisión: Dado que los registros solo abarcan 12 trimestres se evaluó si es que ese valor atípico alteraba significativamente la media de los registros y se encontró que la diferencia era de 1.02 puntos porcentuales. Por ello, se decidió mantener dicho valor con la finalidad de mantener la integridad de la serie temporal de 12 trimestres y evitar introducir valores artificiales.

7.6. Limpieza de la tabla objetivos ESG

- **Estandarización de Texto:** Se realizó una evaluación de coherencia y estandarización de texto en las columnas descriptivas **categoria_esg** y **kpi_nombre** para asegurar que los nombres de las categorías y KPIs sean uniformes para evitar errores al filtrar y agrupar datos en el *dashboard*.

  Decisión: Se confirmó la consistencia en la ortografía y el uso de mayúsculas/minúsculas en los nombres de las categorías (Ambiental, Social, Gobernanza, Financiero) y los nombres de los KPIs (Tasa Reciclaje, Rotación, Margen Neto, etc.).

- **Eliminación de columnas auxiliares:** La columna **valor_ejemplo** fue eliminada debido a que no representaba datos reales medidos ni objetivos oficiales, sino valores referenciales que podrían generar confusión en el análisis.