Gobierno de Datos – Eco-Moda S.A.S.

Este documento establece:

1. Roles y responsabilidades sobre la gestión de los datos
2. Reglas de calidad aplicadas en el pipeline de limpieza ya implementado
3. Estándares de nomenclatura y formato para garantizar consistencia

El gobierno de datos se aplica durante todo el ciclo de vida de los datos, desde su ingesta hasta su consumo en análisis y modelos predictivos. Aunque actualmente se trabaja con datos históricos, las políticas aquí establecidas están diseñadas para ser aplicables también cuando se integren fuentes de datos en tiempo real.

1. DEFINICIÓN DE ROLES Y RESPONSABILIDADES

El gobierno de datos requiere la asignación clara de roles para garantizar la calidad, seguridad y trazabilidad de la información. A continuación, se definen los roles necesarios para Eco-Moda.

1.1 Chief Data Officer (CDO) / Líder de Datos

Perfil sugerido: Director de Transformación Digital de Eco-Moda
Responsabilidades principales:

- Aprobar y actualizar las políticas del gobierno de datos
- Resolver conflictos entre áreas respecto a definiciones y uso de datos
- Priorizar iniciativas de mejora de calidad de datos
- Reportar métricas de calidad y gobierno a la dirección ejecutiva
- Asegurar los recursos necesarios para la implementación del gobierno

Autoridad: Toma decisiones finales sobre estándares de datos y políticas corporativas relacionadas.

1.2 Data Steward (Custodio de Datos)

Los Data Stewards son expertos del negocio responsables de la calidad y el significado de los datos en su dominio específico. No requieren ser técnicos, pero deben tener conocimiento profundo del área de negocio que representan.
Para Eco-Moda se proponen los siguientes Data Stewards:

1.2.1 Data Steward - Área Comercial

Perfil sugerido: Gerente de Ventas o Jefe de Analítica Comercial
Responsabilidades:

- Definir el significado de negocio de las categorías de productos (product_group_name, garment_group_name)
- Establecer los valores válidos para clasificaciones de productos
- Definir rangos de precio aceptables según el portafolio de Eco-Moda
- Aprobar reglas de estandarización de categorías inconsistentes
- Validar la coherencia de datos de ventas y productos

1.2.2 Data Steward - Área de Clientes

Perfil sugerido: Gerente de CRM o Marketing
Responsabilidades:

- Definir segmentos de clientes (club_member_status, fashion_news_frequency)
- Establecer políticas de privacidad para datos.
- Validar la coherencia de preferencias y comportamientos de clientes
- Definir métricas de lealtad y retención

1.3 Data Engineer (Ingeniero de Datos)

Equipo: 2 ingenieros especializados
Responsabilidades:

- Implementar pipelines de ingesta, transformación y limpieza de datos
- Traducir reglas de negocio definidas por Data Stewards a código ejecutable
- Desarrollar y mantener validaciones automáticas de calidad
- Documentar procesos y transformaciones aplicadas
- Configurar alertas cuando se detecten problemas de calidad
- Resolver incidentes técnicos en los pipelines

Herramientas utilizadas: Python, SQL, Pandas, bibliotecas de validación de datos

1.4 Data Analyst / Data Scientist

Equipo: 1 analista o científico de datos
Responsabilidades:

- Consumir datos limpios para análisis exploratorios y modelos predictivos
- Reportar problemas de calidad detectados durante el análisis
- Colaborar con Data Engineers en la definición de características (features) para modelos de ML
- Validar la coherencia de métricas en dashboards y reportes
- Documentar modelos predictivos desarrollados

1.5 Usuarios de Negocio

Perfil: Equipos de Marketing, Operaciones, Finanzas, Logística
Responsabilidades:

- Consumir dashboards y reportes según sus permisos de acceso
- Reportar inconsistencias o dudas sobre los datos
- Respetar políticas de acceso y uso de información sensible
- Solicitar nuevos análisis a través de los Data Stewards

Derechos:

- Acceso a dashboards corporativos (OKR/BSC)
- Consulta del catálogo de datos para entender métricas
- Soporte de analistas para interpretación de resultados

2. REGLAS DE CALIDAD DE DATOS

Las reglas de calidad definen los criterios que los datos deben cumplir para ser considerados válidos y confiables.

2.1 Reglas de Completitud

El dataset presenta una calidad excepcional de completitud, con los campos críticos para el negocio 100% completos. Las reglas de completitud se enfocan en mantener este estándar y gestionar los casos minoritarios de valores faltantes.

2.1.1 Campos Críticos

Los siguientes campos son críticos para el negocio y se validó que están completamente poblados:

Campo	**Justificación de criticidad**
transaction_date	Esencial para análisis temporal y tendencias
price	Define el valor de la transacción
product_group_name	Clasificación principal de productos
age	Variable clave para segmentación demográfica

Política: Estos campos deben mantenerse obligatorios en futuras integraciones de datos. Cualquier registro sin estos valores debe ser rechazado en la ingesta.

2.1.2 Campos Importantes

Además de *product_group_name*, existen otros campos que describen las características del producto vendido y son necesarios para realizar análisis más detallados sobre estilos, colores y líneas de negocio.

Estos incluyen atributos como el tipo de prenda, el patrón gráfico, el grupo de color, y la familia de colección. Aunque no son tan críticos como los campos anteriores, aportan un alto valor analítico para el estudio de tendencias, segmentación por estilo y optimización del portafolio de productos.

Política: Estos campos deben mantenerse obligatorios en la medida de lo posible. En caso de que se detecten valores nulos, no se rechazarán los registros, pero se deberá:

- Imputar el valor como "UNKNOWN".
- Registrar la incidencia en los reportes de calidad de datos.
- Revisar la fuente de origen para prevenir futuras omisiones.

2.1.3 Campos con Valores Faltantes

Los siguientes campos presentan valores nulos en menos del 0.5% de los registros:

Campo	Acción de imputación
club_member_status	Imputar como "UNKNOWN"
fashion_news_frequency	Imputar como "NONE"

Justificación de imputaciones:

- **"UNKNOWN" para membresía:** Representa explícitamente clientes sin estado de membresía registrado, permitiendo análisis diferenciado de este segmento
- **"NONE" para newsletter:** Asume que sin preferencia explícita, el cliente no recibe comunicaciones

2.2 Reglas de Unicidad

Para el dataset, dos registros se considerarían duplicados si tienen valores idénticos en los siguientes campos:

- transaction_date
- product_group_name
- garment_group_name
- price
- age
- club_member_status

Política: En futuras integraciones de datos, se implementará un proceso automático de detección y eliminación de registros duplicados.

2.3 Reglas de Rangos Válidos

Acá se establecen límites lógicos del negocio para variables numéricas.

2.3.1 Rango de Precio

Regla: 0.0006 USD ≤ price ≤ 0.507 USD

Justificación basada en EDA:

- **Límite inferior (0.0006):** Valor mínimo detectado en el dataset real
- **Límite superior (0.507):** Valor máximo detectado en el dataset real
- **Distribución observada:** La mayoría de transacciones (>80%) están concentradas en el rango 0.01-0.10 USD

Acción si no cumple: Rechazar registro y documentar en tabla de excepciones

Nota: Los límites se establecieron basándose en los valores reales observados. Para fases futuras, se recomienda revisar estos límites con el Data Steward Comercial para validar si reflejan el portafolio de productos de Eco-Moda.

2.3.2 Rango de Edad

Regla: 16 ≤ age ≤ 99

Justificación basada en EDA:

- **Límite inferior (16):** Clientes menores de 16 años compran acompañados por adultos y no se registran individualmente
- **Límite superior (99):** Valor máximo detectado en el dataset real
- **Distribución observada:** Mayor concentración en rangos 20-40 años (media: 36.89 años)

Acción si no cumple: Rechazar registro

Nota: Se detectaron registros con edades extremas (99 años). Se recomienda al Data Steward de Clientes validar si estos registros son legítimos o requieren corrección en la fuente.

2.3.3 Rango de Fechas

Regla: transaction_date ≥ 2018-01-01 y transaction_date ≤ fecha actual

Justificación: Garantiza la validez temporal del dato dentro del rango histórico y hasta el momento presente.

Acción si no cumple: Rechazar registros con fechas anteriores a 2018 o futuras respecto a la fecha de carga.

3. ESTÁNDARES DE NOMENCLATURA Y FORMATO

Los estándares garantizan consistencia en la estructura, nombres y formatos de los datos.

3.1 Nomenclatura de Columnas

Convención adoptada: snake_case (minúsculas con guiones bajos)

Principios:

- Usar nombres descriptivos y completos (evitar abreviaturas ambiguas)

- Separar palabras con guion bajo (_)
- Escribir completamente en minúsculas
- No utilizar caracteres especiales ni espacios

Ejemplos correctos:

- transaction_date (fecha de transacción)
- product_group_name (grupo de producto)
- customer_age (edad del cliente)
- club_member_status (estado de membresía)
- price_usd (precio en dólares)

Nota sobre el dataset actual: El campo Transaction_Date mantiene PascalCase por legado del dataset original. En futuras integraciones se estandarizará a transaction_date.

3.2 Formatos de Datos

3.2.1 Fechas

Estándar adoptado: ISO 8601: YYYY-MM-DD

Justificación:

- Estándar internacional reconocido
- Evita ambigüedad entre formatos DD/MM/YYYY y MM/DD/YYYY
- Compatible con ordenamiento cronológico automático
- Facilita operaciones de bases de datos

3.2.2 Valores Monetarios

Estándar adoptado: ISO 4217 (USD como moneda base)

Formato de precio:

- **Moneda base:** USD (dólares americanos)
- **Tipo de dato:** Decimal (float)
- **Separador decimal:** Punto (.)
- **Precisión:** Exactamente 2 decimales

3.2.3 Variables Categóricas (Texto)

Capitalización según tipo:

Tipo de categoría	Formato	Ejemplos
Categorías de negocio (productos, estilos)	Title Case *(Primera Letra Mayúscula)*	"T-Shirt", "Casual Wear", "Garment Upper Body"
Estados	UPPER_CASE *(Todo en mayúsculas)*	"ACTIVE", "PRE-CREATE", "LEFT CLUB"
Colores	Title Case	"Black", "Dark Blue", "Light Grey"
Frecuencias / Preferencias	Title Case	"Regularly", "Monthly", "None"

Valores imputados / desconocidos	UPPER_CASE	"UNKNOWN", "NONE",

Principios:

- **No usar valores vacíos ("")** - siempre usar valores explícitos o NULL
- **Consistencia:** Usar siempre el mismo valor para el mismo significado

4.3 Tipos de Datos

Asignación de tipos según naturaleza del campo:

Tipo de información	Tipo de dato sugerido
Fechas de transacción (transaction_date)	datetime64
Precios / valores monetarios (price)	float64
Categorías de producto (negocio) (product_group_name, garment_group_name)	category
Atributos visuales de producto (graphical_appearance_name, colour_group_name, perceived_colour_value_name, perceived_colour_master_name, index_name, index_group_name)	category
Edad del cliente (age)	int64
Segmentación de clientes (club_member_status, fashion_news_frequency)	category