

# Validación y Limpieza de Datos

## Empaquetamiento de Servicios Fijos 2023-2024

Análisis de Datos con Python

Postdata - Dataset Telecomunicaciones

8 de diciembre de 2025

# Contenido

- 1 Descripción del Dataset
- 2 Actividad 1: Validación de Estructura
- 3 Actividad 2: Análisis de Valores Nulos
- 4 Actividad 3: Consistencia de Datos
- 5 Actividad 4: Verificación de Duplicados
- 6 Decisiones de Limpieza
- 7 Resumen Final

# Información General

## Fuente de Datos

- **Fuente:** Postdata - Empaquetamiento de Servicios Fijos
- **URL:** [www.postdata.gov.co/dataset/empaquetamiento-de-servicios-fijos](http://www.postdata.gov.co/dataset/empaquetamiento-de-servicios-fijos)
- **Formato:** CSV (602.87 MB)

## Dimensiones del Dataset

- **Registros totales:** 2,989,763
- **Registros 2023-2024:** 1,750,659 (filtrados para análisis)
- **Campos:** 21 columnas
- **Rango temporal:** 2022 - 2025 (Q2)

# Distribución por Año

## Registros por Año

- **2023:** 898,318 (51.31 %)
- **2024:** 852,341 (48.69 %)

## Distribución Trimestral

- Q1: 439,737 (25.12 %)
- Q2: 438,224 (25.03 %)
- Q3: 447,378 (25.55 %)
- Q4: 425,320 (24.29 %)

## Observación

Distribución balanceada entre años y trimestres, lo que facilita análisis comparativos.

# 1.1 Validación de Campos

## Resultado

✓ Todos los campos obligatorios están presentes

## Campos del Dataset

- **Temporales:** ANNO, TRIMESTRE
- **Identificadores:** ID\_EMPRESA, ID\_DEPARTAMENTO, ID\_MUNICIPIO, ID\_SEGMENTO, ID\_SERVICIO\_PAQUETE, ID\_TECNOLOGIA\_ACCESO, ID\_ESTADO
- **Descriptivos:** EMPRESA, DEPARTAMENTO, MUNICIPIO, SEGMENTO, SERVICIO\_PAQUETE, TECNOLOGIA, ESTADO
- **Mediciones:** CANTIDAD\_LINEAS\_ACCESOS, VALOR\_FACTURADO\_O\_COBRADO, OTROS\_VALORES\_FACTURADOS, VELOCIDAD\_EFECTIVA\_DOWNSTREAM, VELOCIDAD\_EFECTIVA\_UPSTREAM

## 1.2 Conversión de Tipos de Datos

### Campos Convertidos

Campo	Tipo Original	Tipo Convertido
VALOR_FACTURADO_O_COBRADO	object	int64
OTROS_VALORES_FACTURADOS	object	int64
VELOCIDAD_EFECTIVA_DOWNSTREAM	object	float64
VELOCIDAD_EFECTIVA_UPSTREAM	object	float64

### Decisión

Conversión necesaria para realizar cálculos numéricos y análisis estadístico.

# 1.3 Validación de Rangos

## Rangos Validados

- ✓ **ID\_SERVICIO\_PAQUETE**: Todos en rango 1-7
- ✓ **ID\_ESTADO**: Todos en rango 1-2
- ✓ **TRIMESTRE**: Todos en rango 1-4

## Nota sobre ID\_SEGMENTO

- El diccionario indica valores 1-9
- Los datos reales contienen valores 101-109 y 117
- Representan el mismo concepto (último dígito coincide)
- No se considera error, sino actualización del sistema de codificación

## 1.4 Cobertura del Dataset

### Cobertura Geográfica

- **Departamentos:** 33
- **Municipios:** 1,037
- **Top 3 Departamentos:**
  - Antioquia: 15.36 %
  - Cundinamarca: 13.40 %
  - Valle del Cauca: 9.85 %

### Operadores

- **Total empresas:** 1,553
- **Top 3 Operadores:**
  - Comcel: 32.63 %
  - Claro: 14.99 %
  - UNE EPM: 12.26 %

# 1.5 Distribución por Segmento

ID	Segmento	Registros ( %)
107	Corporativo	453,695 (25.92 %)
102	Residencial - Estrato 2	414,245 (23.66 %)
103	Residencial - Estrato 3	290,140 (16.57 %)
101	Residencial - Estrato 1	280,251 (16.01 %)
104	Residencial - Estrato 4	147,517 (8.43 %)

## Insight

El segmento corporativo y estratos 2-3 representan el 66 % del mercado.

# 1.6 Distribución por Servicio

ID	Servicio	Registros ( % )
1	Internet fijo	735,137 (41.99 %)
7	Triple Play	365,134 (20.86 %)
4	Duo Play 1 (Tel + Internet)	292,895 (16.73 %)
5	Duo Play 2 (Internet + TV)	190,681 (10.89 %)
3	Televisión por suscripción	119,254 (6.81 %)
2	Telefonía fija	35,848 (2.05 %)
6	Duo Play 3 (Tel + TV)	11,710 (0.67 %)

## Insight

El 90 % de los servicios incluyen internet (individual o empaquetado).

# 1.7 Distribución por Tecnología

Tecnología	%
FTTH	32.00 %
HFC	17.44 %
Cable	15.48 %
xDSL	10.38 %
NA	9.53 %
Inalámbricas	5.43 %

## Observaciones

- Predominio de fibra óptica (FTTH: 32 %)
- Tecnologías de cable significativas (HFC, Cable: 33 %)
- NA aplica para servicios sin internet

## 2.1 Identificación de Valores Nulos

Campo	Nulos	%
VELOCIDAD_EFECTIVA_UPSTREAM	69,974	4.00 %
VELOCIDAD_EFECTIVA_DOWNSTREAM	23,768	1.36 %
VALOR_FACTURADO_O_COBRADO	1,354	0.08 %
OTROS_VALORES_FACTURADOS	17	0.00 %

### Observación

Total de nulos: 95,113 (0.31 % del dataset). Bajo impacto general.

## 2.2 Contexto de Nulos en Velocidades

### Análisis por Tipo de Servicio

- **Servicios SIN internet** (Tel, TV): 0 nulos
- **Servicios CON internet:**
  - Internet fijo: 3.17 % (downstream), 7.53 % (upstream)
  - Duo Play 1: 0.16 % (downstream), 3.53 % (upstream)
  - Triple Play: 0.00 % (downstream), 0.26 % (upstream)

### Conclusión

Los nulos en velocidades ocurren principalmente en servicios que SÍ requieren internet, indicando datos faltantes reales que deben imputarse.

### 3.1 Casos de Inconsistencia

Caso	Registros	%
Líneas > 0, Valor = 0	104,364	5.96 %
Líneas = 0, Valor > 0	13,011	0.74 %
Líneas = 0, Valor = 0	2,217	0.13 %
<b>Consistentes</b>	<b>1,629,713</b>	<b>93.09 %</b>

#### Interpretación

- 93 % de consistencia es excelente
- 5.96 % sin facturación: servicios gratuitos o uso interno
- 0.74 % facturados sin líneas: posibles errores de registro

### 3.2 Análisis de Valor por Línea

#### Estadísticas Generales

- Media: \$784,483 COP
- Mediana: \$192,857 COP
- Q1: \$105,042 COP
- Q3: \$313,260 COP

#### Valores Atípicos

- Outliers altos: 178,302 (10.18 %)
- Umbral: > \$625,587
- Máximo: \$2,166M COP

#### Insight

Gran diferencia entre media y mediana indica presencia de valores extremos en segmento corporativo.

### 3.3 Valor por Línea según Servicio

Servicio	Mediana (COP)
Internet fijo	\$225,000
Duo Play 2 (Internet + TV)	\$225,232
Triple Play	\$195,617
Duo Play 1 (Tel + Internet)	\$158,250
Televisión	\$139,687
Duo Play 3 (Tel + TV)	\$110,917
Telefonía fija	\$55,201

#### Observación

Servicios con internet tienen valores más altos. Triple Play no es el más costoso por línea.

## 4.1 Duplicados Completos

### Análisis

- **Registros duplicados:** 900 (0.05 %)
- **Grupos:** 533
- **Promedio por grupo:** 1.69

### Conclusión

Nivel de duplicación completa extremadamente bajo. Excelente calidad del dataset.

## 4.2 Duplicados por Dimensiones de Negocio

### Criterio Evaluado

Municipio + Departamento + Segmento + Servicio + Año-Trimestre

Métrica	Valor
Registros "duplicados"	1,698,761 (97.04 %)
Grupos únicos	111,360
Promedio por grupo	15.25

### Interpretación

No son duplicados reales, sino múltiples registros válidos por combinación.

## 4.3 ¿Qué Diferencia los "Duplicados?

### Campos que Varían

Campo	Valores únicos promedio
VALOR_FACTURADO_O_COBRADO	14.45
VELOCIDAD_EFECTIVA_DOWNSTREAM	7.68
CANTIDAD_LINEAS_ACCESOS	7.09
VELOCIDAD_EFECTIVA_UPSTREAM	6.62
ID_EMPRESA	2.88
ID_TECNOLOGIA_ACCESO	2.02
ID_ESTADO	1.49

## 4.4 Conclusión sobre Duplicados

### Razones de la "Duplicación"

Por cada combinación municipio-segmento-servicio-trimestre existen múltiples registros porque:

- ① Diferentes **empresas** operan en la misma zona
- ② Misma empresa ofrece múltiples **tecnologías**
- ③ Diferentes **planes de velocidad**
- ④ Estados diferentes (activo/suspendido)

### Decisión

**Mantener todos los registros.** Representan desagregación válida del mercado.

## 4.5 Duplicados por Dimensión Individual

Dimensión	Duplicados	Únicos
Municipio-Departamento	100 %	1,122
Segmento	100 %	10
Servicio/Paquete	100 %	7
Año-Trimestre	100 %	8

### Interpretación

Comportamiento esperado: cada dimensión individual se repite en múltiples combinaciones. Confirma la granularidad del dataset.

## 5.1 Corrección del Campo TECNOLOGIA

### Problema Identificado

- Valor inconsistente: "NA (No Aplica)"
- Total de registros afectados: 166,913 (9.53 %)

### Solución Aplicada

- Estandarización: "NA (No Aplica)" → "NA"
- Facilita filtros y análisis posteriores

## 5.2 Estrategia de Imputación - Valores Facturados

### Método

Imputación con **mediana por departamento y servicio**

Campo	Nulos	Imputados	Restantes
VALOR_FACTURADO	1,354	1,354	0
OTROS_VALORES	17	17	0

### Justificación

La mediana por departamento-servicio captura variación regional y características del servicio, siendo más robusta que la media ante outliers.

## 5.3 Estrategia de Imputación - Velocidades

### Lógica Aplicada

① **Servicios SIN internet** (Tel, TV): → 0

② **Servicios CON internet:**

- Primario: Mediana por departamento-servicio
- Fallback: Mediana global del servicio
- Último recurso: 0

Campo	Sin Internet	Con Internet	Restantes
DOWNSTREAM	0 → 0	23,768 → mediana	0
UPSTREAM	0 → 0	69,974 → mediana	0

## 5.4 Justificación de la Imputación

### Ventajas del Método

- **Contexto geográfico:** Refleja diferencias regionales en infraestructura
- **Tipo de servicio:** Respeta características propias de cada paquete
- **Robustez:** Mediana no se ve afectada por outliers
- **Lógica de negocio:** 0 para servicios que no requieren internet

### Alternativas Descartadas

- Eliminar registros: Pérdida innecesaria de información
- Imputación global: Ignora variación regional
- Media: Sensible a valores extremos corporativos

## 5.5 Decisión sobre Duplicados

### Decisión Final

**MANTENER todos los registros "duplicados"**

### Justificación

- ① Representan desagregación válida del mercado
- ② Cada registro aporta información única sobre:
  - Diferentes operadores
  - Múltiples tecnologías de acceso
  - Diversos planes de velocidad
  - Estados de servicio
- ③ Permiten análisis granular de competencia
- ④ Esenciales para análisis de participación de mercado

# Resumen de Resultados - Parte 1

## Dataset Limpio

- **Registros:** 1,750,659
- **Campos:** 22 (incluye campo calculado VALOR\_TOTAL\_FACTURADO)
- **Nulos finales:** 16,582 (campos agregados en proceso)
- **Duplicados completos:** 900 (0.05 %)

## Transformaciones Aplicadas

- Conversión de 4 campos a tipos numéricicos
- Estandarización del campo TECNOLOGIA
- Imputación de 95,113 valores nulos
- Validación de rangos y consistencia

# Calidad del Dataset

## Fortalezas

- 93 % consistencia líneas-valor
- 99.95 % sin duplicados completos
- Cobertura nacional (33 deptos)
- Distribución temporal balanceada

## Áreas de Atención

- 4 % nulos en velocidad upstream
- 6 % servicios sin facturación
- Documentación desactualizada (IDs)
- Valores extremos corporativos

## Conclusión General

Dataset de alta calidad, apto para análisis exploratorio y modelado estadístico.

# Hallazgos Principales

- ① **Dominio del Internet:** 90 % de servicios incluyen internet
- ② **Modernización:** 32 % usa fibra óptica (FTTH)
- ③ **Concentración:** Top 3 operadores controlan 60 % del mercado
- ④ **Segmentación:** Corporativo + E2-E3 = 66 % del mercado
- ⑤ **Precios:** Alta variabilidad (mediana: \$192K, máx: \$2,166M)
- ⑥ **Geografía:** Antioquia, Cundinamarca y Valle = 39 % nacional

# Recomendaciones Técnicas

## Para Análisis Futuro

- ① Segmentar análisis por tipo de cliente (residencial vs corporativo)
- ② Analizar evolución de tecnologías (xDSL vs FTTH)
- ③ Estudiar penetración por estrato socioeconómico
- ④ Investigar brechas regionales en cobertura
- ⑤ Evaluar competencia por zona geográfica

## Próximo Paso

**Parte 2:** Análisis exploratorio detallado con visualizaciones y dashboard en Streamlit

# Gracias

¿Preguntas?