Reporte General de Hallazgos 1.1

1. Introducción

El presente análisis tiene como objetivo examinar los datos del archivo 01_DiatomInventories_GTstudentproject_B.csv para identificar patrones, distribuciones y tendencias en las variables cuantitativas y cualitativas del dataframe. Los hallazgos permitirán comprender mejor las operaciones de muestreo y las características de las columnas registradas.

2. Metodología

Fuente de datos: Archivo CSV proporcionado.

Limpieza de datos:

- Identificación y reemplazo de valores nulos con la mediana.
- Detección de valores atípicos mediante el criterio de desviaciones estándar.

Variables analizadas:

- **Cualitativas:** TaxonName, TaxonCode, SamplingOperations_code, Date_SamplingOperation.
- **Cuantitativas:** Abundance_nbcell, TotalAbundance_SamplingOperation, Abundance_pm.

Herramientas utilizadas: Python (Pandas, NumPy, Matplotlib).

Análisis aplicado:

- Estadística descriptiva (mínimos, máximos, rangos, frecuencias).
- Visualizaciones: diagramas de caja, gráficos de barras y gráficos de área.
- Clasificación de variables cuantitativas en categorías según la regla de Sturges.

3. Hallazgos Principales

3.1 Valores Nulos y Datos Faltantes

Valores nulos iniciales por columna:

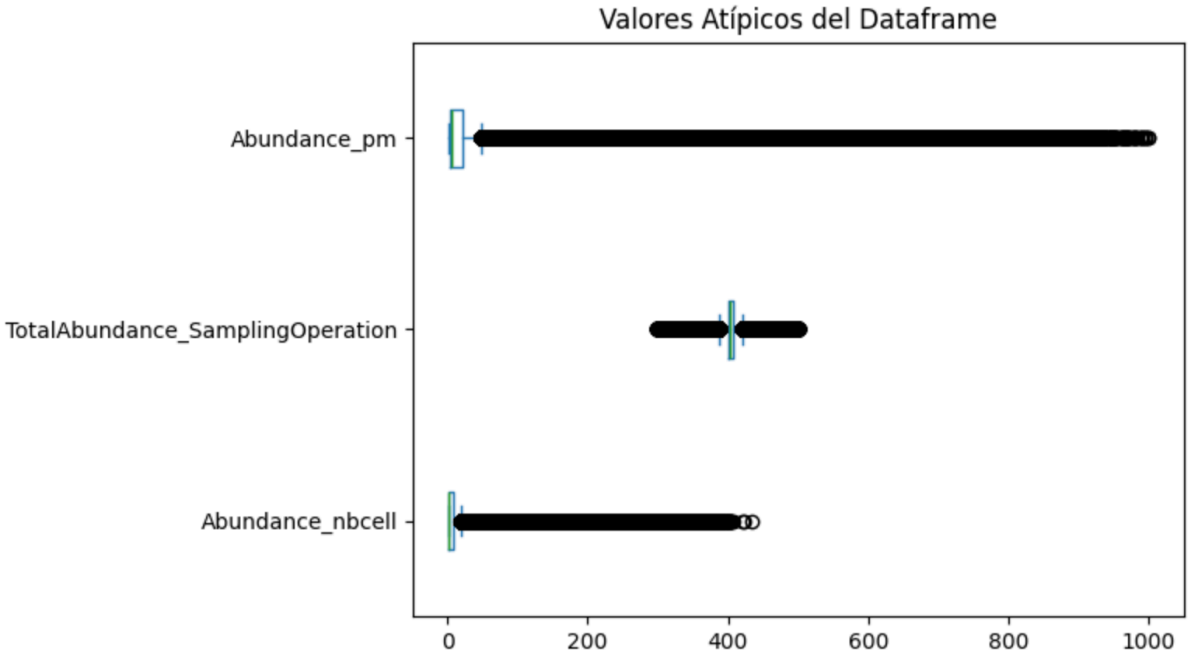
```
Abundance_nbcell                    37619
TotalAbundance_SamplingOperation    34628
Abundance_pm                        37352
dtype: int64
```

Tras la imputación con la mediana, todas las columnas quedaron completas:

```
TaxonName                           0
TaxonCode                           0
SamplingOperations_code             0
CodeSite_SamplingOperations         0
Date_SamplingOperation              0
Abundance_nbcell                    0
TotalAbundance_SamplingOperation    0
Abundance_pm                        0
dtype: int64
```

3.2 Valores Atípicos en Variables Cuantitativas

Diagrama de caja de variables cuantitativas:



Límites superior e inferior:

Limite superior:

```
Abundance_nbcell                    94.948382
TotalAbundance_SamplingOperation   437.096025
Abundance_pm                       233.878730
dtype: float64
```

Limite inferior:

```
Abundance_nbcell                    -71.124054
TotalAbundance_SamplingOperation    374.808697
Abundance_pm                        -175.176372
dtype: float64
```

Se eliminaron los valores fuera de estos límites para asegurar un análisis confiable.

3.3 Distribución de Variables Cualitativas

1. TaxonName(TOP 10):
    a. Tabla de frecuencias de los taxones más representativos

TaxonName	count
Achnanthes minutissima	43691
Amphora pediculus	39209
Cocconeis euglypta	38570
Sellaphora nigri	38039
Navicula cryptotenella	37723
Nitzschia dissipata	34461
Vibrio tripunctatus	30899
Rhoicosphenia abbreviata	30560
Navicula permitis	29789
Achnanthes lanceolata	27239

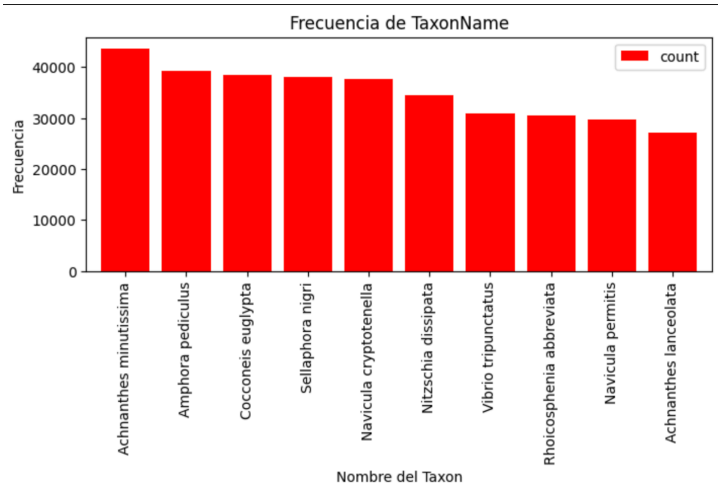
Achnanthes minutissima es el taxón más frecuente, con 43,691 conteos.

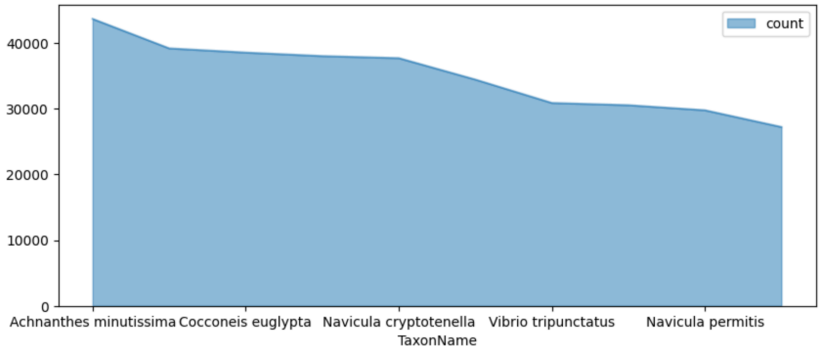
Achnanthes lanceolata es el menos frecuente de los que se muestran, con 27,239 conteos.

El resto de los taxones se encuentran entre estos dos valores.

Esta tabla permite una visualización de la frecuencia de cada taxón, facilitando la comparación directa de los valores.

    b. gráficos:





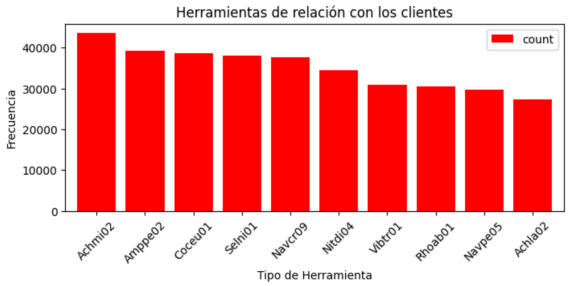
Tanto la tabla como los gráficos indican que Achnanthes minutissima es el taxón más frecuentemente observado, mientras que el taxón menos frecuente de los mostrados es Achnanthes lanceolata. El patrón muestra una clara dominancia de Achnanthes minutissima, con el resto de los taxones principales siguiendo un patrón de disminución gradual.

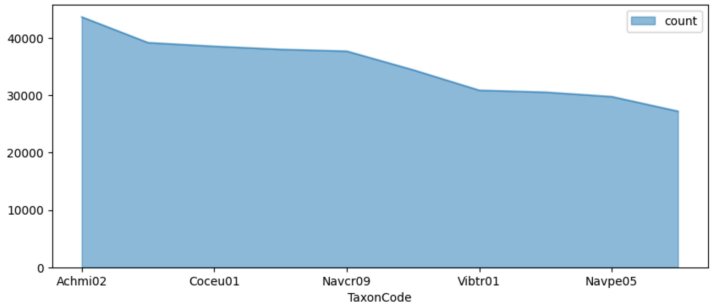
2. TaxonCode:
    a. Tabla de frecuencias:

TaxonCode	count
Achmi02	43691
Amppe02	39209
Coceu01	38570
Selni01	38039
Navcr09	37723
Nitdi04	34461
Vibtr01	30899
Rhoab01	30560
Navpe05	29789
Achla02	27239

La tabla de frecuencias muestra una lista de códigos de taxones y el número de veces que cada uno se ha registrado, lo que indica su abundancia o representatividad.

b. Gráficos:



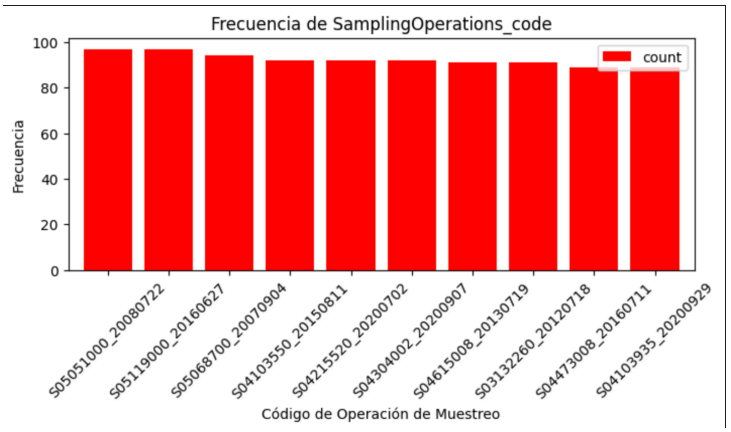


La tabla y los gráficos muestran que el código de taxón **Achmi02** es el más abundante, seguido por Amp pe02 y Coce u01, mientras que el menos común es **Achla02**. Los gráficos ofrecen una representación visual clara de las tendencias y las diferencias de frecuencia, complementando la precisión de los datos numéricos de la tabla. El patrón muestra una clara dominancia del **TaxonCode Achmi02** en los registros, con el resto de los códigos principales siguiendo un patrón de disminución gradual y constante en su abundancia.

3. SamplingOperations_code:
   a. Tabla de frecuencias:

SamplingOperations_code	count
S05051000_20080722	97
S05119000_20160627	97
S05068700_20070904	94
S04103550_20150811	92
S04215520_20200702	92
S04304002_20200907	92
S04615008_20130719	91
S03132260_20120718	91
S04473008_20160711	89
S04103935_20200929	89

b. Gráfico de barras:

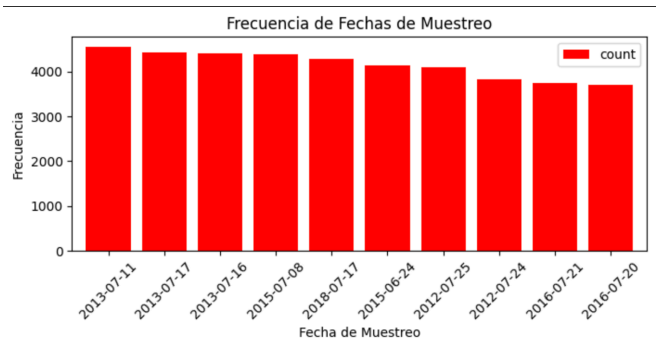


La tabla y el gráfico muestran una distribución de frecuencias con poca variabilidad para los códigos de muestreo, La tabla da los valores exactos y el gráfico permite una comparación visual rápida. El patrón muestra una alta consistencia en la generación de datos entre las diferentes operaciones, lo que sugiere un proceso de recolección de datos uniforme.

4. Date_SamplingOperation:
   a. Tabla de frecuencias:

Date_SamplingOperation	count
2013-07-11	4545
2013-07-17	4431
2013-07-16	4401
2015-07-08	4376
2018-07-17	4278
2015-06-24	4126
2012-07-25	4091
2012-07-24	3833
2016-07-21	3742
2016-07-20	3706

   b. Gráfico de barras



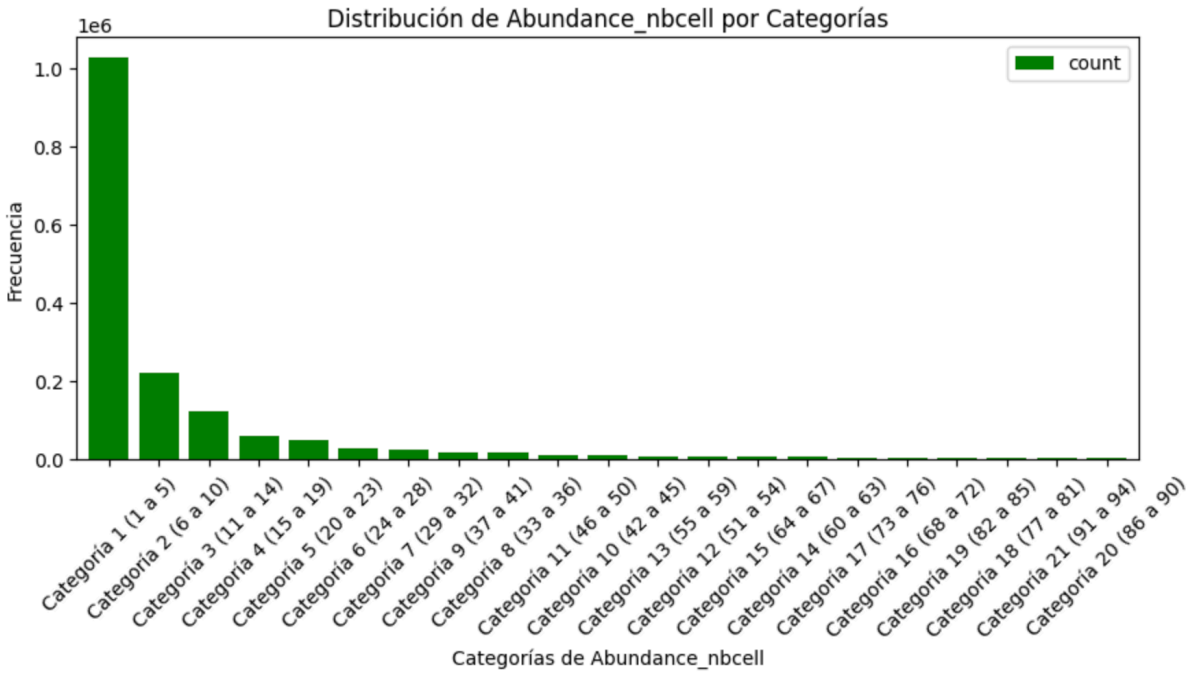
La tabla como el gráfico de barras muestran que la fecha 2013-07-11 tuvo la mayor cantidad de operaciones de muestreo, mientras que la fecha 2016-07-20 tuvo la menor la gráfica reduce de forma gradual lo que significa un declive.

3.4 Distribución de Variables Cuantitativas por Categorías

1. Abundance_nbcell:
    a. Tabla de frecuencias por categorías:

Abundance_nbcell	count
Categoría 1 (1 a 5)	1029011
Categoría 2 (6 a 10)	222430
Categoría 3 (11 a 14)	124121
Categoría 4 (15 a 19)	58398
Categoría 5 (20 a 23)	48843
Categoría 6 (24 a 28)	27953
Categoría 7 (29 a 32)	25819
Categoría 9 (37 a 41)	16873
Categoría 8 (33 a 36)	16490
Categoría 11 (46 a 50)	11287
Categoría 10 (42 a 45)	10917
Categoría 13 (55 a 59)	8411
Categoría 12 (51 a 54)	7798
Categoría 15 (64 a 67)	6317
Categoría 14 (60 a 63)	5875
Categoría 17 (73 a 76)	4985
Categoría 16 (68 a 72)	4446
Categoría 19 (82 a 85)	3995
Categoría 18 (77 a 81)	3687
Categoría 21 (91 a 94)	3303
Categoría 20 (86 a 90)	2913

    b. Gráfico de barras:

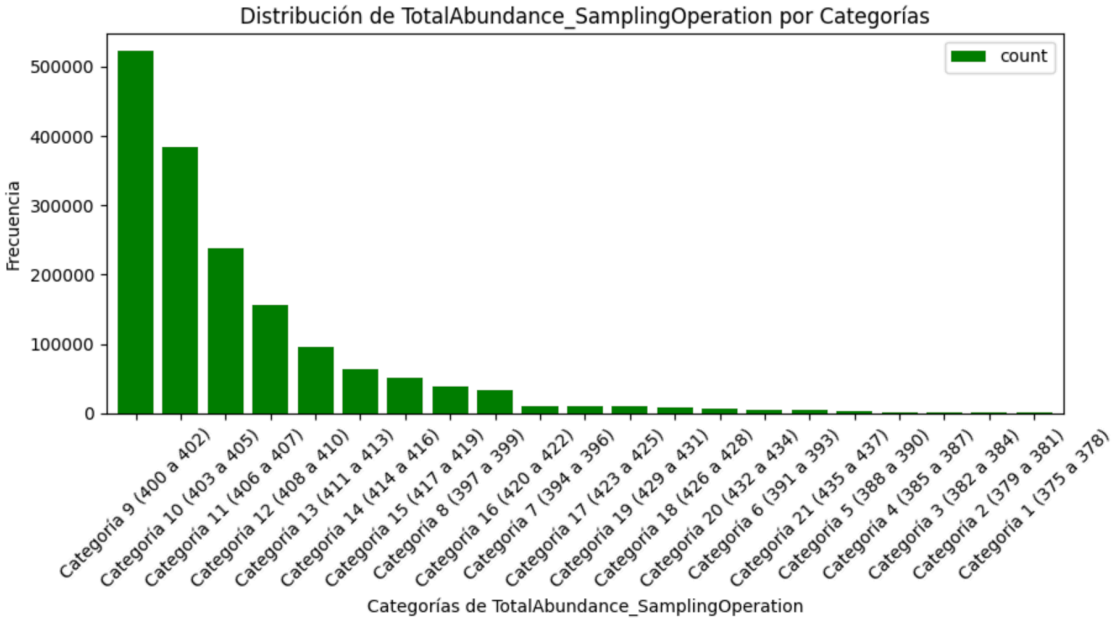


La tabla y el gráfico demuestran que los valores de Abundance_nbcell con menor abundancia (1 a 5) son extremadamente comunes, mientras que los valores de mayor abundancia son considerablemente raros. El gráfico es particularmente efectivo para ilustrar esta fuerte asimetría en la distribución de los datos, categorías cayendo de manera exponencial a medida que aumenta el número de células.

2. TotalAbundance_SamplingOperation:
    a. Tabla de frecuencias por categorías:

TotalAbundance_SamplingOperation	count
Categoría 9 (400 a 402)	522234
Categoría 10 (403 a 405)	383742
Categoría 11 (406 a 407)	237322
Categoría 12 (408 a 410)	155486
Categoría 13 (411 a 413)	95422
Categoría 14 (414 a 416)	63255
Categoría 15 (417 a 419)	51264
Categoría 8 (397 a 399)	38516
Categoría 16 (420 a 422)	32822
Categoría 7 (394 a 396)	10984
Categoría 17 (423 a 425)	10959
Categoría 19 (429 a 431)	9527
Categoría 18 (426 a 428)	9001
Categoría 20 (432 a 434)	6069
Categoría 6 (391 a 393)	5514
Categoría 21 (435 a 437)	4428
Categoría 5 (388 a 390)	2775
Categoría 4 (385 a 387)	1540
Categoría 3 (382 a 384)	1350
Categoría 2 (379 a 381)	1012
Categoría 1 (375 a 378)	650

b. Gráfico de barras:

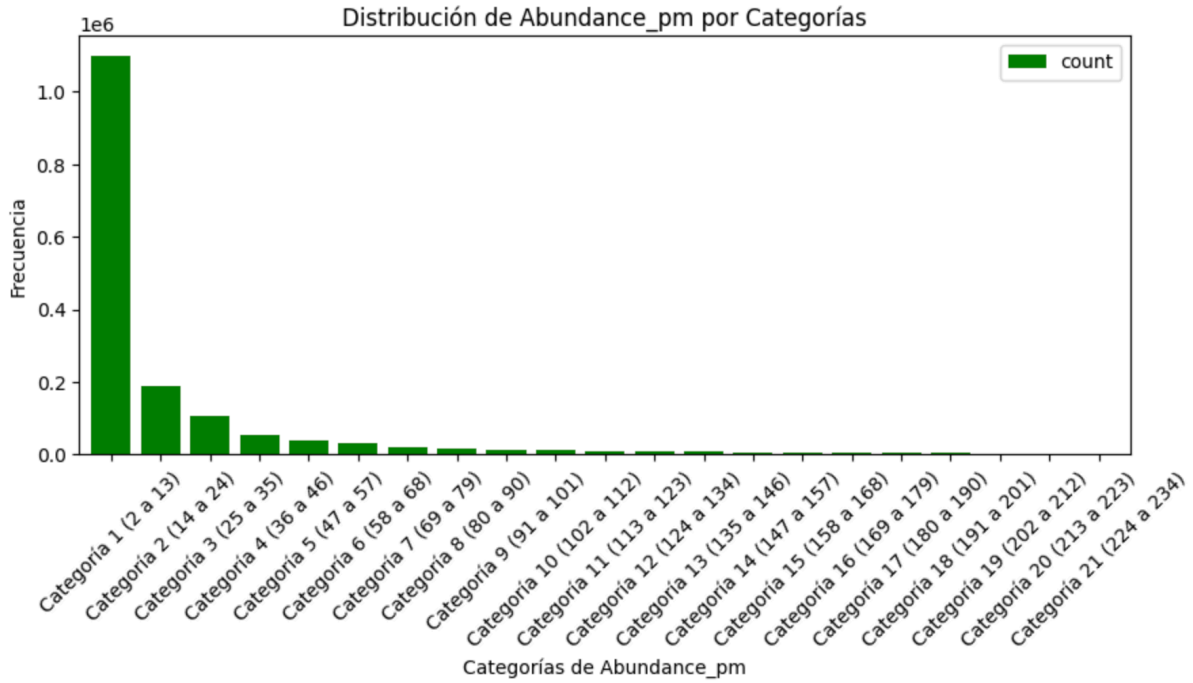


La tabla y el gráfico indican que la mayoría de las operaciones de muestreo tienen una abundancia total que se sitúa en el rango de 400 a 402. Las frecuencias caen rápidamente en otras categorías, lo que sugiere una distribución de datos muy concentrada.

3. Abundance_pm:
   a. Tabla de frecuencias por categorías:

Abundance_pm	count
Categoría 1 (2 a 13)	1097633
Categoría 2 (14 a 24)	187224
Categoría 3 (25 a 35)	107706
Categoría 4 (36 a 46)	54461
Categoría 5 (47 a 57)	39998
Categoría 6 (58 a 68)	30440
Categoría 7 (69 a 79)	21536
Categoría 8 (80 a 90)	18417
Categoría 9 (91 a 101)	13935
Categoría 10 (102 a 112)	11741
Categoría 11 (113 a 123)	9975
Categoría 12 (124 a 134)	8474
Categoría 13 (135 a 146)	7467
Categoría 14 (147 a 157)	6261
Categoría 15 (158 a 168)	5761
Categoría 16 (169 a 179)	4753
Categoría 17 (180 a 190)	4374
Categoría 18 (191 a 201)	4069
Categoría 19 (202 a 212)	3476
Categoría 20 (213 a 223)	3320
Categoría 21 (224 a 234)	2851

b. Gráfico de barras:



La tabla y el gráfico de barras muestran que los valores de Abundance_pm con menor abundancia (rango de 2 a 13) son abrumadoramente más comunes, mientras que los valores de abundancia más altos son muy raros. El gráfico es

particularmente útil para ilustrar esta fuerte concentración de datos en una sola categoría.

5. Conclusiones

El análisis permitió identificar una fuerte concentración de datos en categorías bajas de abundancia, con escasa representatividad de valores altos. Asimismo, se observó que pocos taxones y códigos dominan la distribución, mientras que la mayoría aparece con menor frecuencia. La limpieza de nulos y atípicos aseguró mayor confiabilidad en los resultados, y los patrones encontrados facilitan la comprensión de la dinámica de muestreo y de la representatividad de los taxones registrados.