Tecnológico de Monterrey

Analítica de datos y herramientas de inteligencia artificial II

Grupo 101

Regresión Lineal Simple y Múltiple

DataForge

Jesús Eduardo Valle Villegas | A01770616

Manuel Eduardo Covarrubias Rodríguez | A01737781

Diego Antonio Oropeza Linarte | A01733018

Ithandehui Joselyn Espinoza Mazón | A01734547

Mauricio Grau Gutierrez Rubio | A01734914

Última edición el 30 de Septiembre del 2025

1. Objetivo

El propósito de esta actividad fue aplicar técnicas de regresión lineal simple y múltiple al dataset del Datathon, mediante un preprocesamiento adecuado de las variables categóricas y cuantitativas. El análisis buscó identificar relaciones significativas entre las variables, comparar coeficientes y generar hallazgos que permitan una mejor interpretación de los datos.

2. Metodología

Contexto del análisis

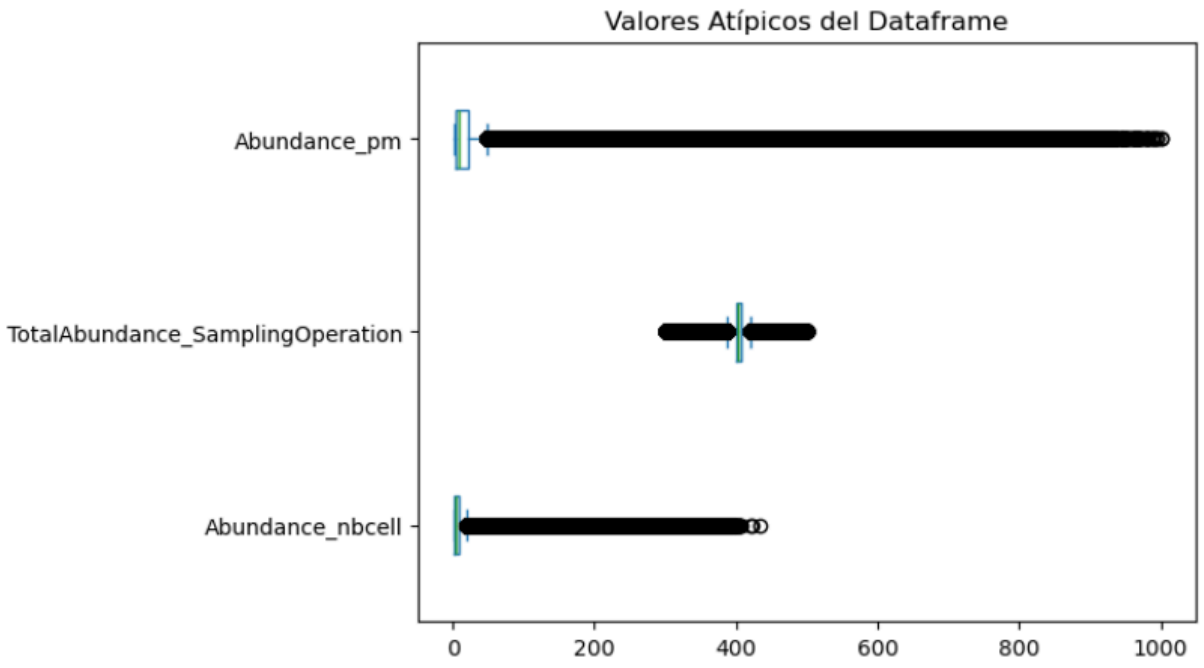
Este análisis se realizó sobre el dataset 01_DiatomInventories_GTstudentproject_B.csv, el cual contiene información proveniente del Datathon relacionada con inventarios de diatomeas y operaciones de muestreo. El objetivo principal fue aplicar técnicas de regresión lineal simple y múltiple para identificar relaciones significativas entre variables, así como comparar coeficientes obtenidos en modelos simples frente a modelos múltiples.

- **Preprocesamiento:** Se transformaron las variables categóricas a numéricas con base en la jerarquía de frecuencias, de modo que las categorías más frecuentes recibieron los valores más bajos. Además, se identificaron y trataron los valores atípicos presentes en el dataframe para garantizar la calidad del análisis. Posteriormente, se validó que el dataframe resultante contuviera únicamente variables numéricas.
- **Regresión lineal simple:** Se generó un heatmap para visualizar la fuerza de las relaciones lineales entre las variables y se seleccionaron los 5 pares con mayor correlación. Se construyeron modelos simples entre los pares más correlacionados con el fin de analizar la dirección e intensidad de sus relaciones.
- **Regresión lineal múltiple :** Se desarrollaron modelos múltiples tomando como dependientes cada una de las variables cuantitativas. Los coeficientes obtenidos fueron comparados con los de los modelos simples y con los valores observados en el mapa de calor.

3. Resultados

3.1 Preprocesamiento

Valores Atípicos:



Interpretación:
El análisis evidenció la presencia de valores atípicos en Abundance_pm y Abundance_nbcell, con una alta concentración en valores bajos pero con casos aislados muy elevados que generan colas largas en la distribución. Esto sugiere que, aunque la mayoría de los registros presentan abundancias reducidas, existen observaciones extremas que podrían influir en los resultados de los modelos.

En contraste, TotalAbundance_SamplingOperation mostró una distribución más compacta, aunque también con algunos valores fuera del rango esperado. La detección de estos casos es relevante, ya que pueden afectar la estabilidad y precisión de las regresiones lineales.

Transformación de variables:

Index	Taxon Name_ num	TaxonCod e_num	SamplingOperations _code_num	CodeSite_SamplingOpe rations_num	Date_SamplingOper ation_num
0	1	1	1	1	1
1	1	1	2	2	2
2	2	2	2	3	3
3	2	2	3	4	4

4	2	2	4	5	5
5	2	2	5	6	6
6	2	2	6	7	7
7	2	2	7	8	8
8	2	2	8	9	9
9	2	2	9	10	10

Con el fin de aplicar técnicas de regresión lineal, fue necesario transformar las variables categóricas en variables numéricas. Para ello, se utilizó la jerarquía de frecuencias, asignando valores más bajos a las categorías con mayor frecuencia de aparición.

La tabla muestra el resultado de esta transformación para las variables *TaxonName*, *TaxonCode*, *SamplingOperations_code*, *CodeSite_SamplingOperations* y *Date_SamplingOperation*. De esta forma, se generaron nuevas columnas numéricas (*_num*) que permiten el tratamiento estadístico y la construcción de modelos de regresión.

3.2 Regresión simple

Top 5 variables con mayor correlación

Ranking	Variable 1	Variable 2	Correlación Original	Interpretación
1	TaxonName_num	TaxonCode_num	1.0000	Muy fuerte
2	Abundance_nbcell	Abundance_pm	0.9890	Muy fuerte
3	SamplingOperations_code_num	CodeSite_SamplingOperations_num	0.3836	Débil
4	CodeSite_SamplingOperations_num	Date_SamplingOperation_num	0.1269	Muy débil

5	**SamplingOperations_code_num**	**Date_SamplingOperation_num**	**0.1207**	**Muy débil**

TaxonName_num ↔ TaxonCode_num

- Correlación: 1.0000
- Interpretación: Muy fuerte
- Relación positiva: cuando *TaxonName_num* aumenta, *TaxonCode_num* también aumenta.

Abundance_nbcell ↔ Abundance_pm

- Correlación: 0.9890
- Interpretación: Muy fuerte
- Relación positiva: un incremento en *Abundance_nbcell* implica un aumento en *Abundance_pm*.

SamplingOperations_code_num ↔ CodeSite_SamplingOperations_num

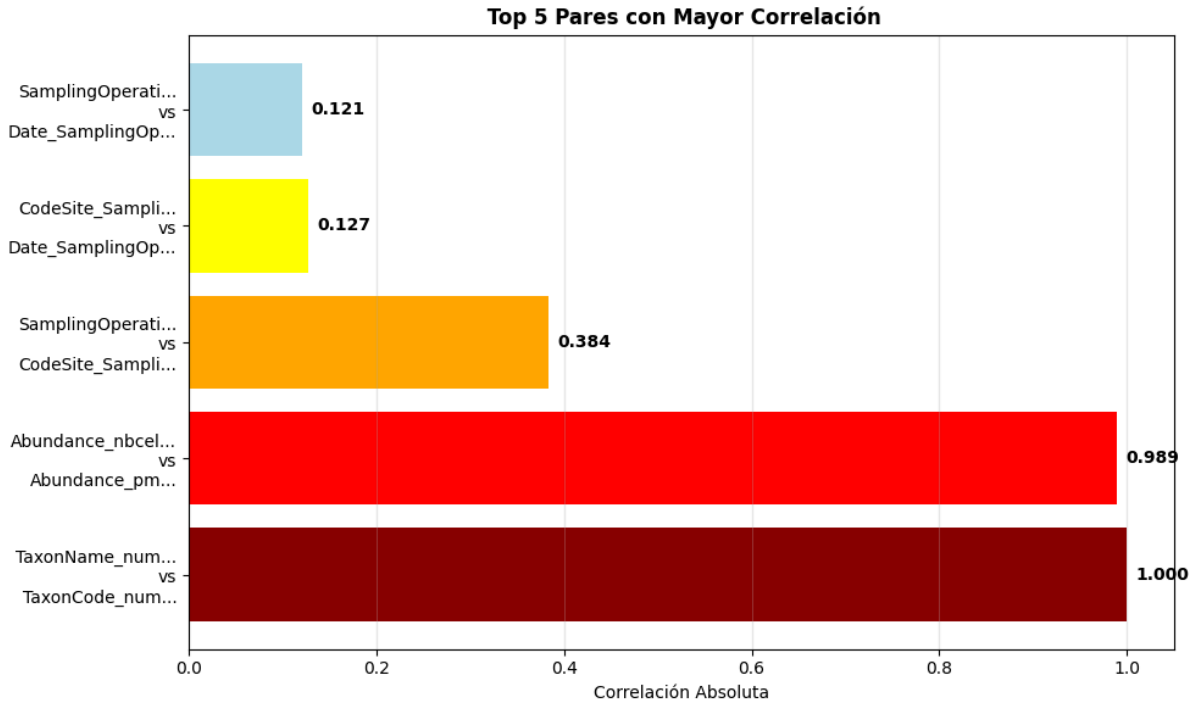
- Correlación: 0.3836
- Interpretación: Débil
- Relación positiva: existe una ligera tendencia conjunta entre ambas variables.

CodeSite_SamplingOperations_num ↔ Date_SamplingOperation_num

- Correlación: 0.1269
- Interpretación: Muy débil
- Relación positiva: la relación es prácticamente nula.

SamplingOperations_code_num ↔ Date_SamplingOperation_num

- Correlación: 0.1207
- Interpretación: Muy débil
- Relación positiva: no se observa un patrón relevante.



Síntesis de hallazgos

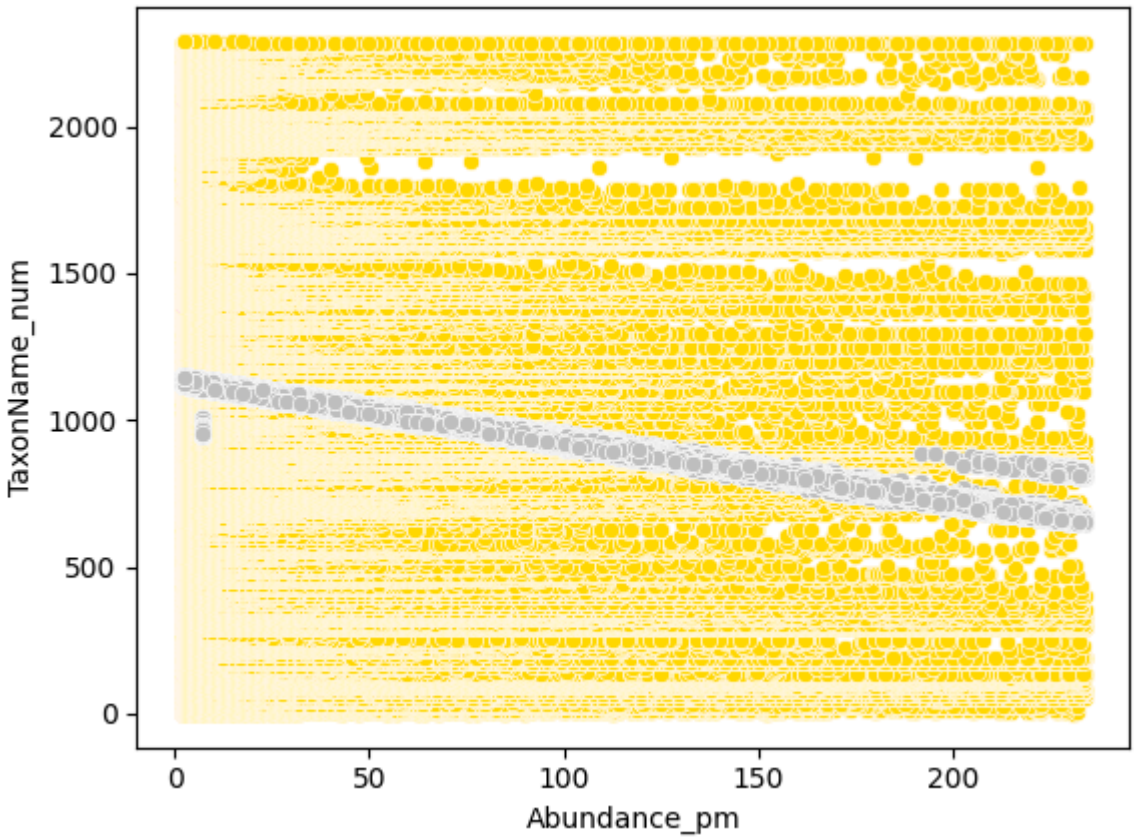
En general, el análisis revela que:

- Solo dos pares de variables presentan correlaciones muy fuertes y estadísticamente significativas.
- El resto de las asociaciones son débiles o muy débiles, lo que limita su poder explicativo en modelos predictivos.
- Estos resultados anticipan que únicamente ciertas variables aportarán valor en modelos de regresión lineal, mientras que otras tendrán un impacto marginal.

3.3 Regresión Múltiple

Taxon Name

Variables independientes: Abundance_nbcell, Abundance_pm, SamplingOperations_code_num
Variable dependiente: TaxonName_num



En el gráfico se observa la relación entre TaxonName_num y Abundance_pm, donde los puntos amarillos representan los valores reales y la línea gris corresponde a la predicción del modelo.

Interpretación de resultados:

1. **Tendencia general**

    - La pendiente negativa de la línea indica una relación inversa entre *Abundance_pm* y *TaxonName_num*, consistente con lo observado en la regresión simple.

    - Sin embargo, el modelo múltiple incorpora más variables independientes, lo cual permite que la predicción sea más robusta y

menos sensible a valores aislados.

2. **Dispersión de los datos**

    - Se observa una amplia dispersión en los valores reales (puntos amarillos), lo que refleja alta variabilidad en los datos.

    - A pesar de esta dispersión, la línea de regresión múltiple mantiene un patrón claro de tendencia, lo que indica que las variables adicionales ayudan a mejorar el ajuste.

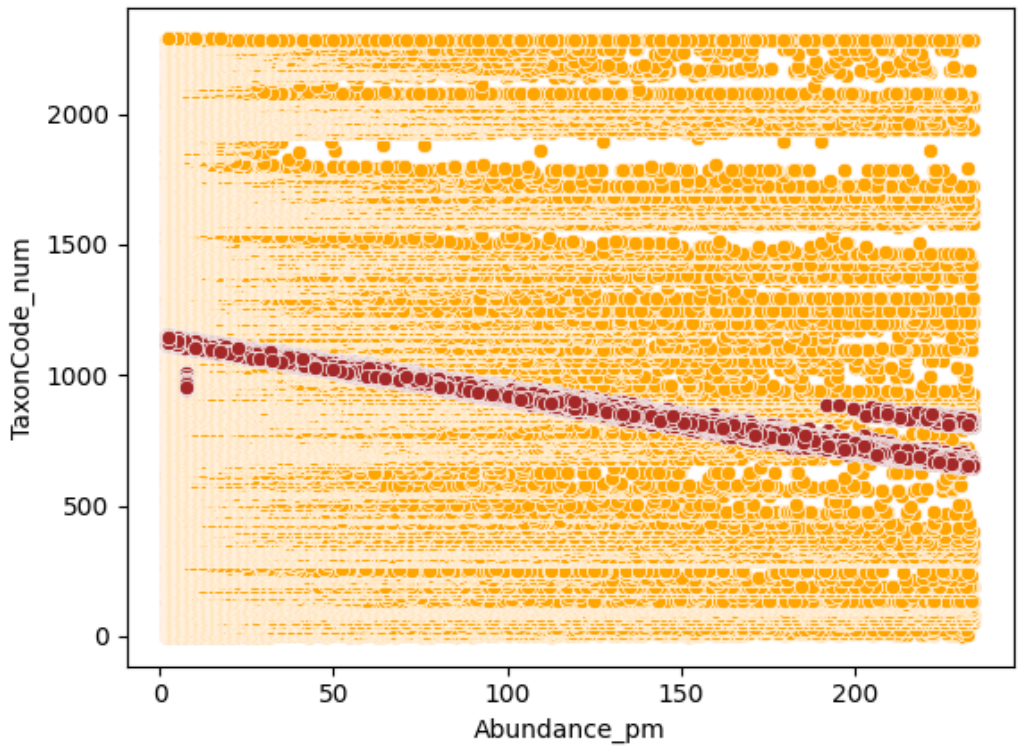
3. **Comparación con regresión simple**

    - Mientras que la regresión simple solo reflejaba la relación entre dos variables, la múltiple permitió capturar de manera más completa la interacción entre *TaxonName_num* y otras variables predictoras.

    - Esto refuerza que los modelos múltiples son más adecuados para datasets complejos con alta variabilidad.

Taxon Code

Variables independientes: Abundance_nbcell, Abundance_pm, SamplingOperations_code_num
Variable dependiente: TaxonCode_num



Relación inversa moderada

- La pendiente de la línea de regresión es negativa, lo que indica una relación inversa: cuando aumenta la abundancia por unidad de medida (*Abundance_pm*), el valor de *TaxonCode_num* tiende a disminuir.

- La pendiente es más marcada que en otras regresiones, lo que sugiere un ajuste más consistente.

Dispersión de los datos

- Los puntos reales presentan alta dispersión en todos los niveles de abundancia, aunque la tendencia descendente sigue siendo visible.

- Esto refleja que, aunque existe una asociación, no es completamente lineal y hay otros factores que influyen en el comportamiento de los taxones.

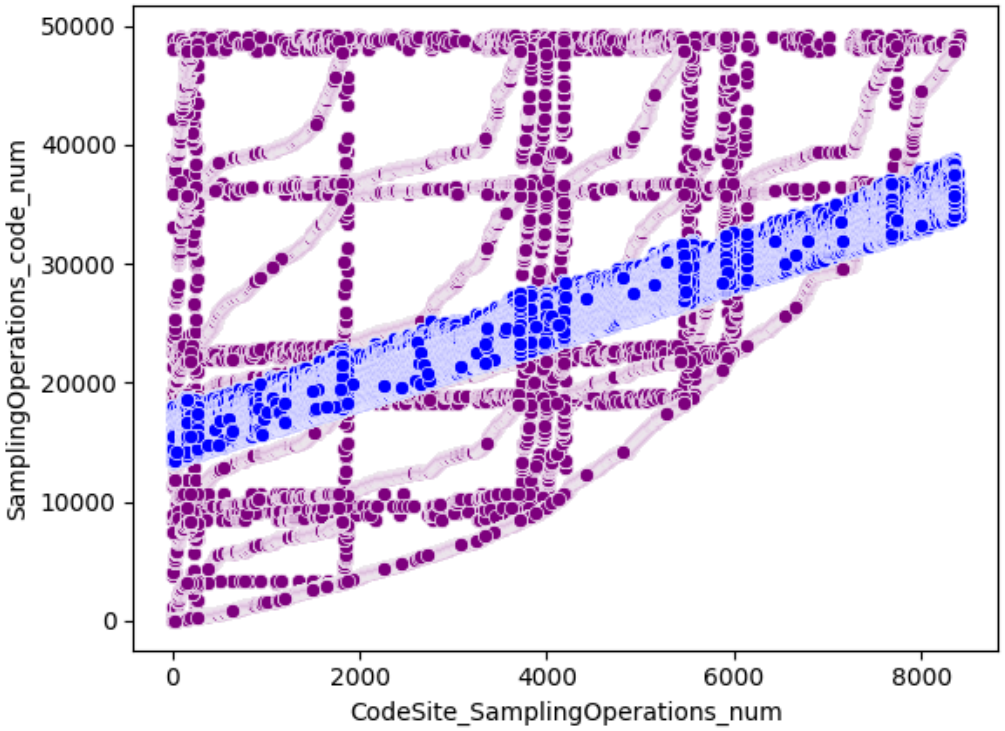
Capacidad del modelo

- A pesar de la dispersión, el modelo logra capturar un patrón claro y consistente en la relación entre ambas variables.

- Esto lo convierte en un ejemplo de regresión simple con mayor relevancia en comparación con otros pares analizados.

SamplingOperations_code

Variables independientes: CodeSite_SamplingOperations_num, Date_SamplingOperation_num
Variable dependiente: SamplingOperations_code_num



La gráfica compara los valores reales (en color **morado**) con los valores predichos por el modelo (en color **azul**).

Interpretación:

1. **Tendencia general**

    - Los valores predichos (azul) siguen una trayectoria lineal ascendente, lo que indica una relación positiva: a medida que aumentan los códigos de sitio de muestreo (*CodeSite_SamplingOperations_num*), también lo hacen los códigos de operaciones de muestreo (*SamplingOperations_code_num*).

    - Esto coincide con la correlación débil positiva encontrada en el análisis previo (≈0.38).

2. **Comparación entre reales y predichos**

    - Los valores reales (morado) presentan una gran dispersión y múltiples patrones en forma de bandas, lo que refleja que no existe una relación lineal clara entre estas variables.

- A pesar de esta dispersión, las predicciones del modelo se concentran en un rango intermedio y muestran un patrón más definido, ajustándose a la tendencia central de los datos.

3. **Capacidad del modelo**

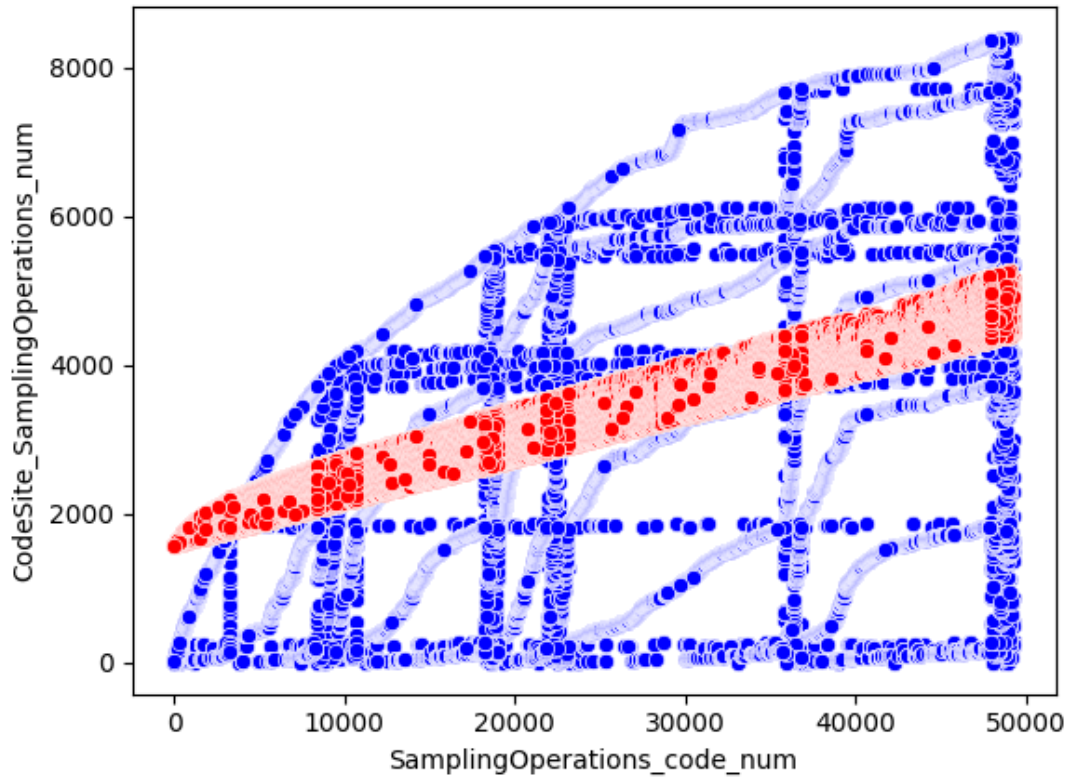
   - Aunque la regresión múltiple logra identificar la dirección de la relación (positiva), el alto grado de dispersión en los datos reales limita la precisión del ajuste.

   - Esto evidencia que la relación entre estas variables no es predominantemente lineal y que podrían intervenir otros factores externos no capturados por el modelo.

CodeSite_SamplingOperations

Variable dependiente: CodeSite_SamplingOperations_num

Variables independientes: SamplingOperations_code_num, Date_SamplingOperation_num



La gráfica compara los valores reales (en color **azul**) con los valores predichos por el modelo (en color **rojo**).

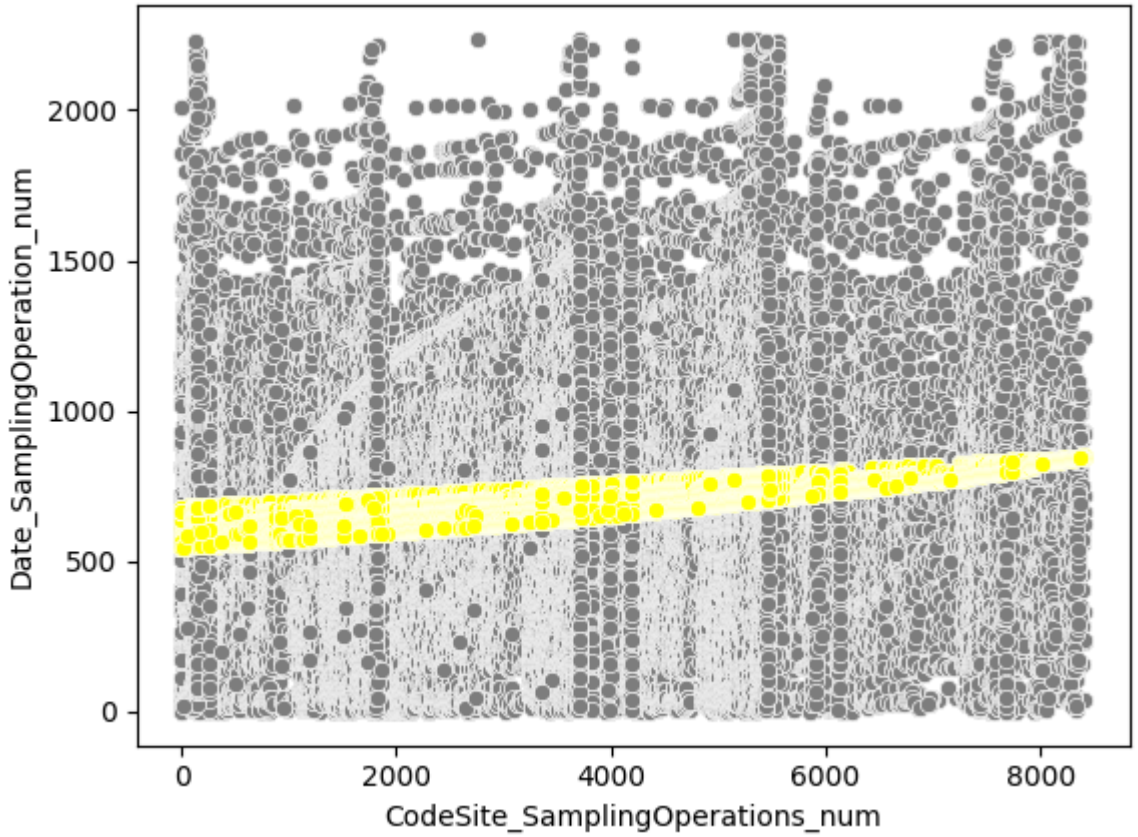
Interpretación:

1. **Relación positiva**
    - La recta de predicciones en rojo indica una relación lineal positiva entre *SamplingOperations_code_num* y *CodeSite_SamplingOperations_num*.
    - A medida que aumentan los códigos de operaciones de muestreo, también tienden a incrementarse los códigos de sitio de muestreo.
2. **Dispersión de datos reales**
    - Los valores reales (azul) presentan un patrón escalonado y muy disperso, lo que refleja que la relación entre ambas variables no sigue estrictamente una línea recta.
    - Esto genera bandas o bloques de valores que dificultan el ajuste exacto del modelo.
3. **Capacidad predictiva del modelo**
    - A pesar de la dispersión, la regresión múltiple logra identificar una tendencia central ascendente.
    - Las predicciones (rojo) se alinean con la dirección general de los datos, aunque no capturan la complejidad de las agrupaciones presentes en los valores reales.

Date_SamplingOperation

Variable dependiente: Date_SamplingOperation_num

Variables independientes: SamplingOperations_code_num, CodeSite_SamplingOperations_num



Relación positiva débil

- La pendiente de las predicciones (amarillo) es levemente ascendente, lo que indica una relación **positiva pero muy débil** entre el sitio de muestreo y la fecha de muestreo.

- Esto coincide con la correlación obtenida en el análisis preliminar (≈0.12).

Dispersión de los datos

- Los valores reales (gris) presentan una dispersión muy alta, con gran variabilidad en las fechas asociadas a diferentes sitios de muestreo.

- La nube de puntos no refleja un patrón lineal fuerte, lo que dificulta que el modelo capture relaciones significativas.

Capacidad del modelo

- El modelo logra identificar una ligera tendencia positiva en la relación, pero las predicciones se concentran en un rango muy limitado en comparación con

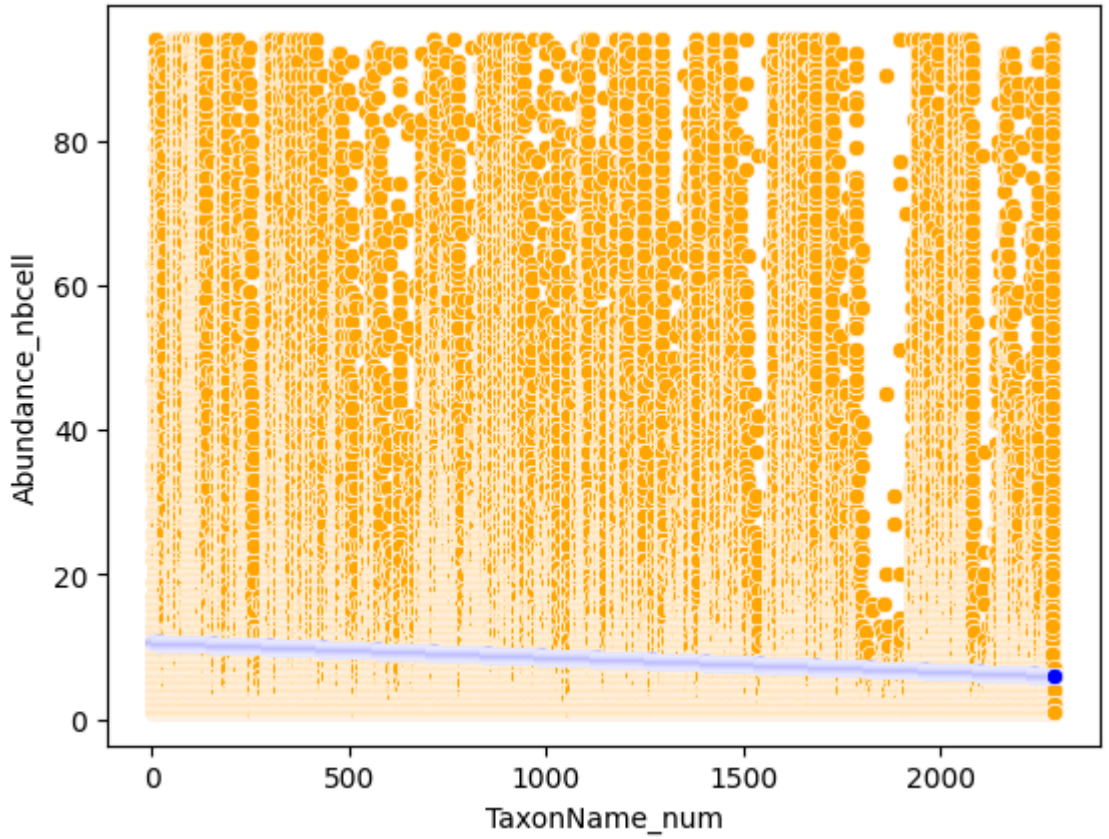
la dispersión real de los datos.

- Esto evidencia que el modelo tiene un poder explicativo muy reducido para esta combinación de variables.

Abundance_nbcell

Variable dependiente: Abundance_nbcell

Variables independientes: TaxonCode_num, TaxonName_num



Relación inversa débil

- La pendiente de la recta es negativa, lo que indica una relación inversa entre ambas variables: a medida que aumenta *TaxonName_num*, los valores de *Abundance_nbcell* tienden a disminuir ligeramente.

- Sin embargo, la pendiente es muy poco pronunciada, lo que refleja que la relación es débil.

Dispersión de los datos

- Los datos reales presentan una distribución muy dispersa, con una alta concentración en valores bajos de abundancia (<20).

- Esto sugiere que la mayoría de los taxones presentan abundancias reducidas, mientras que casos de abundancia elevada son menos frecuentes.

Capacidad explicativa del modelo

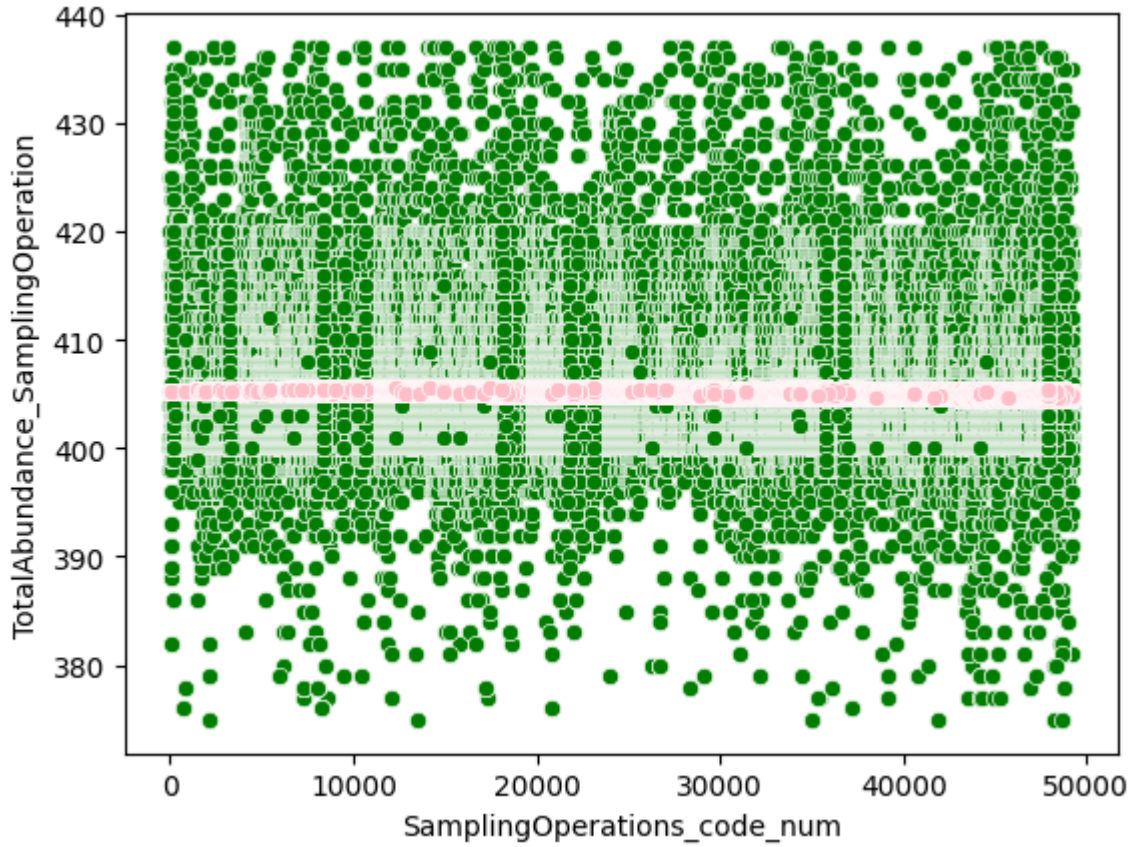
- Debido a la alta dispersión y al predominio de valores bajos, la regresión no logra capturar un patrón fuerte.

- El ajuste lineal solo representa una tendencia general, pero no explica la variabilidad observada en los datos.

TotalAbundance_SamplingOperation

Variable dependiente: TotalAbundance_SamplingOperation

Variables independientes: SamplingOperations_code_num, CodeSite_SamplingOperations_num, Date_SamplingOperation_num



Relación casi nula

- La línea de regresión es prácticamente horizontal, lo que indica una correlación muy débil entre *SamplingOperations_code_num* y *TotalAbundance_SamplingOperation*.

- Esto significa que el código de la operación de muestreo no influye significativamente en la abundancia total observada.

Dispersión de datos

- Los puntos verdes están altamente dispersos alrededor de la recta, sin un patrón lineal claro.

- La mayoría de los valores se concentran alrededor del rango 400 ± 20, lo que refleja una estabilidad relativa en la variable dependiente.

Capacidad del modelo

- El modelo no logra explicar la variabilidad de los datos, ya que prácticamente todos los valores de abundancia total se mantienen dentro de un rango

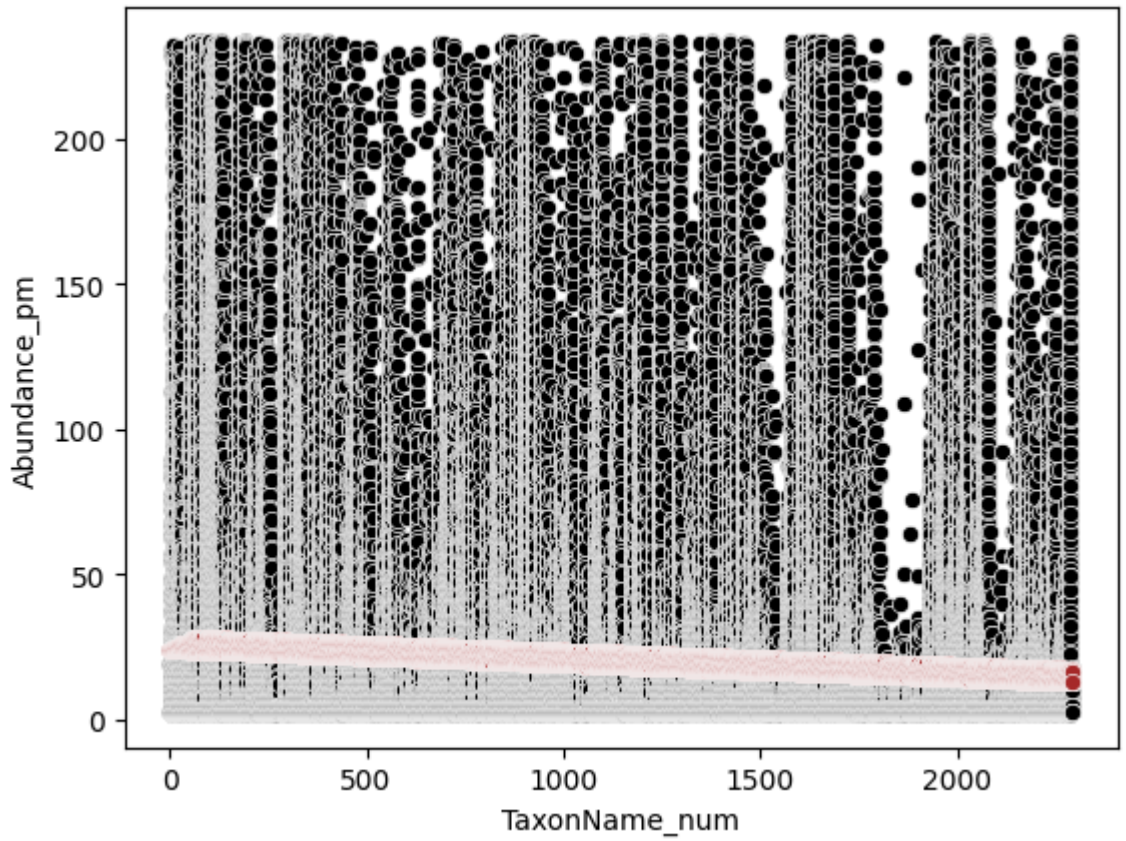
similar, independientemente del código de operación.

- Esto sugiere que esta variable independiente (*SamplingOperations_code_num*) no es un predictor adecuado de la abundancia total.

Abundance_pm

Variables independientes: TaxonName_num, TaxonCode_num, SamplingOperations_code_num

Variable dependiente: Abundance_pm



Relación inversa débil

- La pendiente de la recta es negativa, lo que indica una relación inversa: a medida que aumenta *TaxonName_num*, los valores de *Abundance_pm* tienden a disminuir ligeramente.

- Sin embargo, la inclinación de la recta es muy baja, lo que refleja que la relación es débil y no **explica bien la variabilidad de los datos.**

Alta dispersión de datos

- Los valores reales están altamente dispersos y no siguen un patrón lineal evidente.
- Existen rangos con gran concentración de puntos alrededor de valores bajos de abundancia (<50), lo que genera una distribución heterogénea.

Capacidad del modelo

- El modelo de regresión capta solo una tendencia general descendente, pero no logra capturar la complejidad ni las variaciones de los datos.
- Esto limita su utilidad predictiva para explicar la relación entre taxones y abundancia por unidad de medida.

Resultados de la regresión múltiple:

Variable dependiente	**Variables independientes**	**Coeficientes**	**R^2**
TaxonName_num	Abundance_nbcell, Abundance_pm, SamplingOperations_code_num	Antes:-0.101 Después:0.1016	0.0103
TaxonCode_num	Abundance_nbcell, Abundance_pm, SamplingOperations_code_num	Antes:-0.101 Después:0.1016	0.0103
SamplingOperations_code_num	CodeSite_SamplingOperations_num, Date_SamplingOperation_num	Antes:0.3836 Después:0.3903	0.1524

CodeSite_SamplingOperations_num	SamplingOperations_code_num, Date_SamplingOperation_num	Antes:0.3836 Después:0.3920	0.1537
Date_SamplingOperation_num	SamplingOperations_code_num, CodeSite_SamplingOperations_num	Antes:0.127 Despues 0.149	0.0222
TotalAbundance_SamplingOperation	SamplingOperations_code_num, CodeSite_SamplingOperations_num, Date_SamplingOperation_num	Antes:-0.18 Después:0.029	0.0008
Abundance_nbcell	TaxonCode_num, TaxonName_num	Antes:-0.100 Después:0.104	0.0109
Abundance_pm	TaxonName_num, TaxonCode_num, SamplingOperations_code_num	Antes:-0.101 Después:0.108	0.0116

El incremento del R^2 tras incluir múltiples variables fue marginal (<0.02 en la mayoría de los casos), lo que indica que las variables adicionales no aportaron información predictiva sustancial”
haría el cierre más contundente.

Conclusión

El análisis realizado permitió aplicar técnicas de regresión lineal simple y múltiple sobre el dataset de inventarios de diatomeas, con el objetivo de identificar relaciones entre variables categóricas y cuantitativas. A partir del preprocesamiento se transformaron variables a formato numérico y se detectaron valores atípicos que influyeron en la distribución de los datos.

Los modelos de regresión simple mostraron que solo algunos pares de variables presentan correlaciones muy fuertes (por ejemplo, *TaxonName_num* ↔ *TaxonCode_num* y *Abundance_nbcell* ↔ *Abundance_pm*), mientras que la mayoría de las asociaciones fueron débiles o muy débiles. Esto anticipaba que los modelos múltiples tendrían un poder explicativo limitado.

En los modelos de regresión múltiple se confirmó este comportamiento: aunque la inclusión de más variables independientes permitió capturar tendencias más amplias, los valores de R^2 fueron bajos en casi todos los casos y los incrementos frente a la regresión simple fueron marginales. Además, la dispersión de los datos y la presencia de valores extremos redujeron la capacidad predictiva de los modelos, particularmente en las variables relacionadas con la abundancia.