Reporte — Actividad 2.1 (Regresión Lineal Simple y Múltiple)

Integrantes:

Ivanna Maldonado Cervantes

Paula Simonetta Madrid Pérez

Ania Díaz González

Miranda Eugenia Colorado Arróniz

1. Introducción

El objetivo de este trabajo es aplicar técnicas de regresión lineal (simple y múltiple) sobre el dataset 01_DiatomInventories_GTstudentproject_B.csv que cuenta con los datos del socio formador, de tal manera logramos evaluar relaciones entre variables cuantitativas y categóricas (convertidas a numéricas por frecuencia) y presentar los modelos y su interpretación.

2. Realizado

1. Limpieza:
    a. Eliminación de outliers
    b. Imputación de medianas
    c. Conversión de variables categóricas a numéricas por orden de frecuencia.
2. Análisis realizado:
    a. Estadística descriptiva
    b. Detección de outliers (boxplot)
    c. Matriz de correlación y heatmap
    d. Regresiones múltiples para varias variables dependientes

3. Resultados

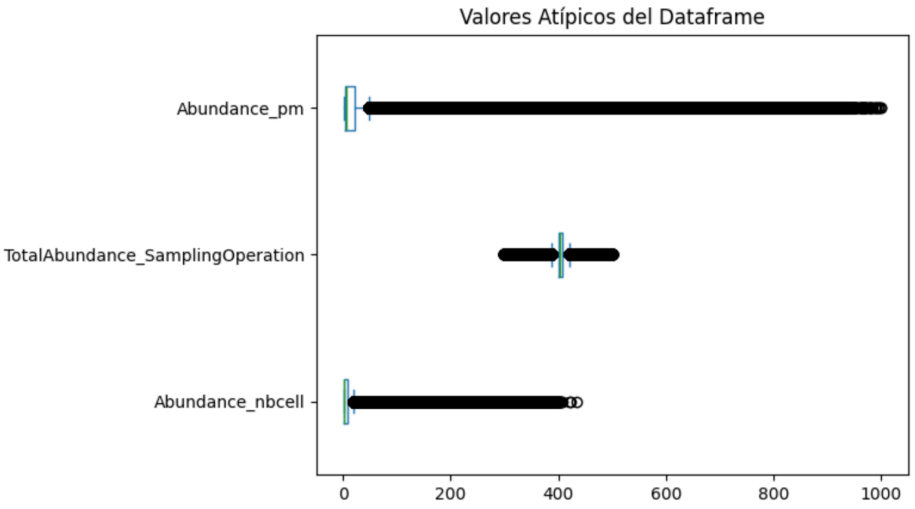
3.1 Estado y limpieza del DataFrame

Info del dataframe (tipos y no nulos) después de limpieza:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1643872 entries, 0 to 1643871
Data columns (total 8 columns):
 #   Column                            Non-Null Count    Dtype
---  ------                            --------------    -----
 0   TaxonName                         1643872 non-null  object
 1   TaxonCode                         1643872 non-null  object
 2   SamplingOperations_code           1643872 non-null  object
 3   CodeSite_SamplingOperations       1643872 non-null  object
 4   Date_SamplingOperation            1643872 non-null  object
 5   Abundance_nbcell                  1643872 non-null  float64
 6   TotalAbundance_SamplingOperation  1643872 non-null  float64
 7   Abundance_pm                      1643872 non-null  float64
dtypes: float64(3), object(5)
memory usage: 100.3+ MB
```

3.2 Detección de outliers (boxplot)

- Diagrama de caja (variables cuantitativas):



El diagrama de caja de variables cuantitativas mostró valores extremos en las abundancias (por ejemplo, Abundance_nbcell y TotalAbundance_SamplingOperation). Estos valores se recortaron a 3 desviaciones estándar para asegurar análisis confiables.

3.3 Frecuencias de variables categóricas (antes de mapear)

- frecuencias:

Se calcularon frecuencias para TaxonName, TaxonCode, SamplingOperations_code, CodeSite_SamplingOperations y Date_SamplingOperation.

– TaxonName / TaxonCode: unas pocas especies y códigos concentran la mayoría de registros, lo que indica que el dataset está dominado por unos pocos taxones.

```
Frecuencias de TaxonName:
TaxonName
Achnanthes minutissima      43691
Amphora pediculus           39209
Cocconeis euglypta          38570
Sellaphora nigri            38039
Navicula cryptotenella      37723
                            ...
Encyonopsis neoamphioxys        1
Encyonopsis recta               1
Lindavia bodanica               1
Leptocylindrus minimus          1
Eunotia perpusilla              1
Name: count, Length: 2292, dtype: int64
```

```
Frecuencias de TaxonCode:
TaxonCode
Achmi02    43691
Amppe02    39209
Coceu01    38570
Selni01    38039
Navcr09    37723
           ...
Encne03        1
Encre01        1
Linbo01        1
Lepmi01        1
Eunpe02        1
Name: count, Length: 2292, dtype: int64
```

– SamplingOperations_code / CodeSite_SamplingOperations: los códigos de operación y de sitio presentan poca variabilidad comparados con los taxones, reflejando operaciones repetidas en pocos sitios.

```
Frecuencias de CodeSite_SamplingOperations:
CodeSite_SamplingOperations
S05119000    864
S05021650    845
S05093300    834
S05021500    831
S05018800    809
            ...
S05183300      8
S04360004      7
S06068415      7
S06113320      4
S06210790      3
Name: count, Length: 8404, dtype: int64
```

```
Frecuencias de SamplingOperations_code:
SamplingOperations_code
S05051000_20080722    97
S05119000_20160627    97
S05068700_20070904    94
S04103550_20150811    92
S04215520_20200702    92
                      ..
S05192040_20170914     2
S05224100_20080821     2
S05221600_20080826     1
S04022000_20150605     1
S05206750_20080826     1
Name: count, Length: 49231, dtype: int64
```

– Date_SamplingOperation: hay fechas con muchos muestreos (picos) y otras con muy pocos, lo que genera alta concentración temporal.

```
Frecuencias de Date_SamplingOperation:
Date_SamplingOperation
2013-07-11    4545
2013-07-17    4431
2013-07-16    4401
2015-07-08    4376
2018-07-17    4278
              ...
2015-02-09       9
2015-02-12       9
2014-02-24       8
2017-02-06       8
2013-06-15       5
Name: count, Length: 2237, dtype: int64
```

```python
df_numeric = df.copy()

variables_categoricas = ['TaxonName', 'TaxonCode', 'SamplingOperations_code',
                         'CodeSite_SamplingOperations', 'Date_SamplingOperation']

for variable in variables_categoricas:
    frecuencias = df[variable].value_counts() #Contamos las frecuencias

    mapeo = {}
    for i, valor in enumerate(frecuencias.index): #Mapeamos el valor de la más frecuente a 1, y asi sucesivamente
        mapeo[valor] = i + 1
    df_numeric[variable] = df[variable].map(mapeo) #Aplicamos el mapeo a la columna
```

✓ 0.4s

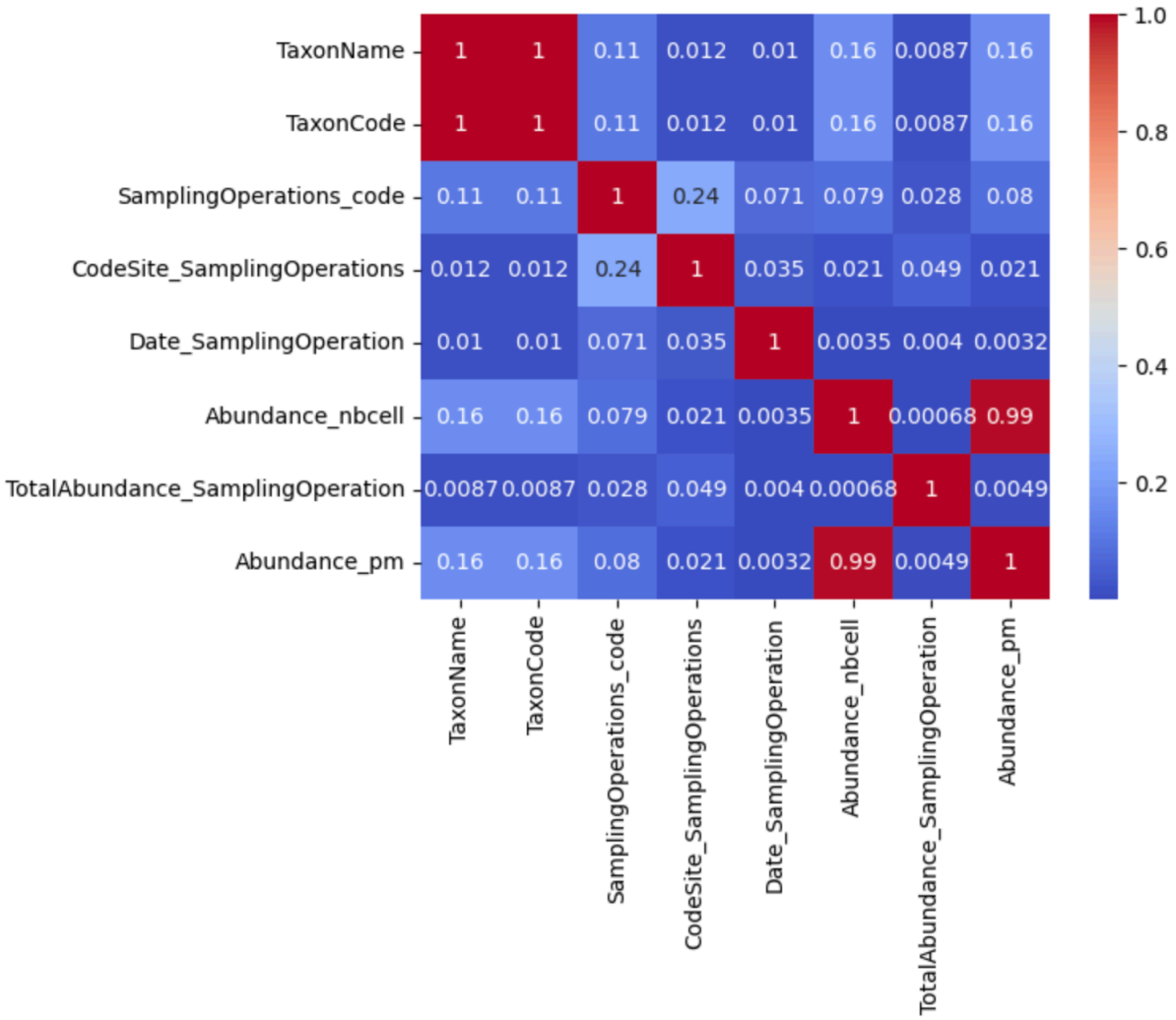
convertimos las variables

3.4 Matriz de correlación y heatmap

- Matriz de correlación (numérica):

	TaxonName	TaxonCode	SamplingOperations_code	CodeSite_SamplingOperations	Date_SamplingOperation	Abundance_nbcell
TaxonName	1.000000	1.000000	-0.107305	-0.011925	0.010478	-0.161495
TaxonCode	1.000000	1.000000	-0.107305	-0.011925	0.010478	-0.161495
SamplingOperations_code	-0.107305	-0.107305	1.000000	0.240210	0.071392	0.079177
deSite_SamplingOperations	-0.011925	-0.011925	0.240210	1.000000	0.034921	0.020863
Date_SamplingOperation	0.010478	0.010478	0.071392	0.034921	1.000000	0.003472
Abundance_nbcell	-0.161495	-0.161495	0.079177	0.020863	0.003472	1.000000
ndance_SamplingOperation	0.008697	0.008697	-0.028288	0.048809	0.004006	0.000679
Abundance_pm	-0.161564	-0.161564	0.079608	0.020519	0.003199	0.989024

- Heatmap (imagen):

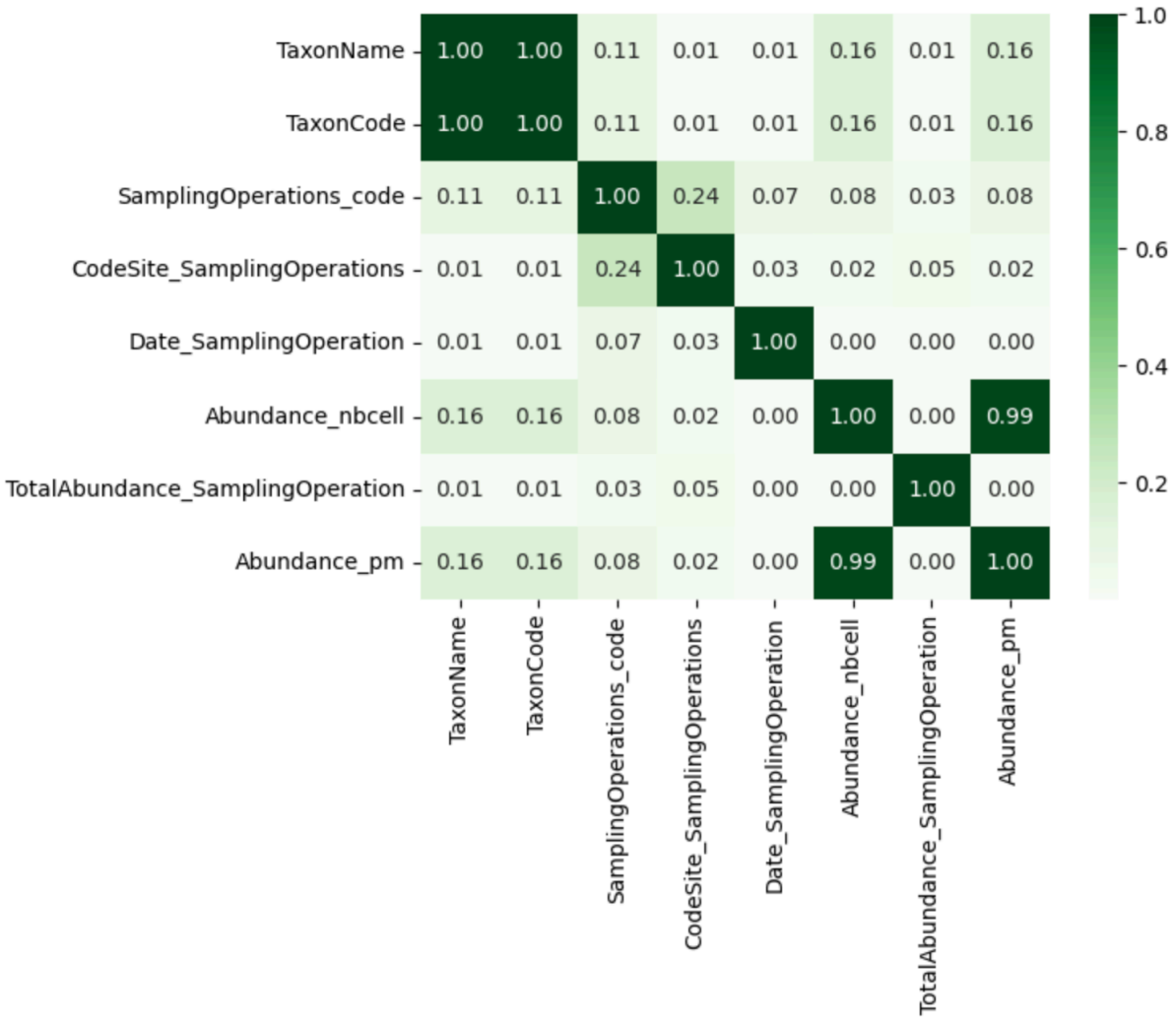


Valores Cercanos a 1 (Rojo Oscuro): Indican una correlación positiva fuerte. Cuando una variable aumenta, la otra tiende a aumentar.

Valores Cercanos a -1 (Azul Oscuro): Indican una correlación negativa fuerte. Cuando una variable aumenta, la otra tiende a disminuir.

Valores Cercanos a 0 (Colores Claros/Medios): Indican una correlación débil. No hay una relación lineal clara entre las variables.

Diagonal Principal (Rojo Intenso con valor 1): Representa la correlación de una variable consigo misma, que siempre es 1.



Los valores en el mapa varían de 0.00 a 1.00 (en este caso, solo se muestran valores absolutos o correlaciones positivas, representadas por tonos de verde más oscuro para correlaciones más fuertes).

Hay colinealidad perfecta entre algunas medidas de abundancia:

- Abundance_nbcell con Abundance_pm
- TotalAbundance_SamplingOperation con Abundance_pm

  Esto indica que miden la misma señal y pueden sustituirse entre sí en modelos.

- Top 5 pares más correlacionados:

	Par de Variables	Correlación	Interpretación
0	TaxonName y TaxonCode	1.00	Correlación perfecta positiva, Ambas variables...
1	Abundance_nbcell y Abundance_pm	1.00	Correlación perfecta positiva, Ambas miden abu...
2	TotalAbundance_SamplingOperation y Abundance_pm	1.00	Correlación perfecta positiva, La abundancia t...
3	TotalAbundance_SamplingOperation y Abundance_n...	1.00	Correlación perfecta positiva, Misma relación ...
4	SamplingOperations_code y CodeSite_SamplingOpe...	0.24	Correlación baja, Muestra una relación débil e...

La tabla lo resume y muestra las correlaciones más fuertes (las primeras 4 = 1.00). Esto alerta sobre la multicolinealidad en los modelos.

Conclusión

Hay fuertes relaciones entre medidas de abundancia; elegir una sola simplifica el modelo.

Los modelos lineales básicos ofrecen predicciones pero, en la mayoría de casos, muestran R^2 bajos: conviene explorar modelos no lineales, más variables o codificaciones distintas.