REGRESIÓN LINEAL SIMPLE Y MÚLTIPLE

DATAFORGE

Analitica de datos y herramientas de inteligencia artificial II

TEAM MEMBERS (DATAFORGE)

JESÚS EDUARDO VALLE VILLEGAS
Finanzas
A01770616

DIEGO ANTONIO OROPEZA LINARTE
BGB
A01733018

MANUEL EDUARDO COVARRUBIAS RODRÍGUEZ
ITC
A01737781

ITHANDEHUI JOSELYN ESPINOZA
ITC
A01734547

MAURICIO GRAU GUTIERREZ RUBIO
LEM
A01734914

ACTIVIDAD 2.1

Objetivo

Aplicar técnicas de regresión lineal simple y múltiple al dataset del Datathon / projectos_forvia.csv, realizando un preprocesamiento adecuado de las variables categóricas y cuantitativas, para identificar relaciones significativas entre variables y comparar los coeficientes obtenidos, con el fin de generar hallazgos que apoyen la interpretación de los datos.

Metodología

Se utilizó un dataset del Datathon para aplicar regresión lineal, con el objetivo de identificar relaciones entre variables y comparar el desempeño de modelos simples y múltiples.

Preprocesamiento

Conversión de variables categóricas a numéricas (frecuencias).

Regresión Simple

Generación de un heatmap para visualizar relaciones lineales y selección de los 5 pares de variables con mayor correlación.

Construcción de modelos entre las variables más correlacionadas para analizar la fuerza y dirección de la relación.

Regresión Múltiple

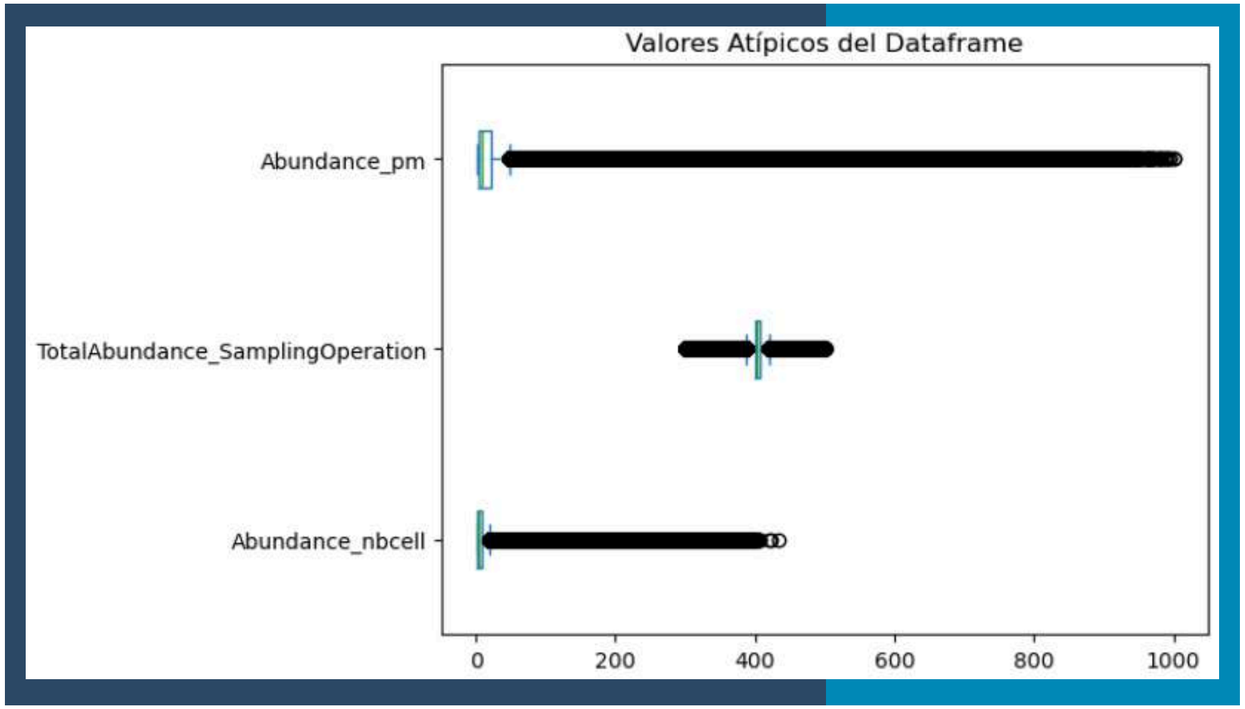
Desarrollo de modelos con varias variables independientes para cada variable cuantitativa, comparando sus coeficientes con los de la regresión simple.

Hallazgos

Identificación de las variables con mayor impacto, análisis de diferencias entre modelos simples y múltiples, y validación de patrones observados en los datos.

Preprocesamiento

- Se codificaron las variables categóricas en valores numéricos según su frecuencia.
- Se detectaron y ajustaron los valores atípicos.
- Se validó que el dataframe final contuviera únicamente variables numéricas.



Transformacion de variables

Mapeo con un ciclo for

Index	TaxonName_num	TaxonCode_num	SamplingOperations_code_num	CodeSite_SamplingOperations_num	Date_SamplingOperation_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

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.

ANÁLISIS DE CORRELACIONES USANDO HEATMAP

Hallazgos

- **TaxonName_num ↔ TaxonCode_num** **Correlación:** 1.0000 (|1.0000|)

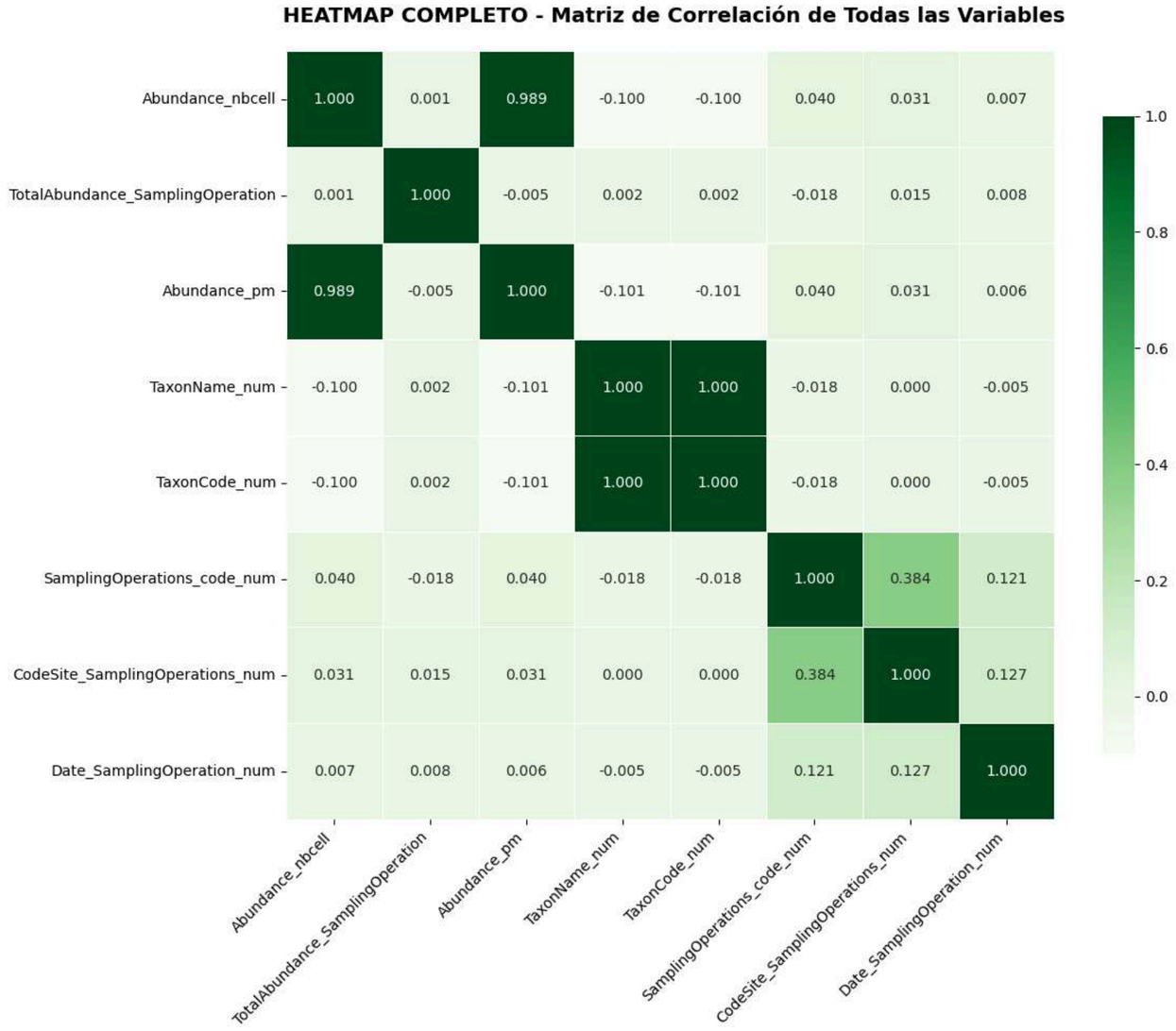
Interpretación: Muy Fuerte Relación **POSITIVA**:

- **Abundance_nbcell ↔ Abundance_pm** **Correlación:** 0.9890 (|0.9890|)

Interpretación: Muy Fuerte Relación **POSITIVA.**

- **SamplingOperations_code_num ↔ CodeSite_SamplingOperations_num** **Correlación:** 0.3836 (|0.3836|)

Interpretación: Débil Relación **POSITIVA**:



ANÁLISIS DE INSIGHTS

Cada diatomea está correctamente identificada y codificada

Hay guerra ecológica entre especies por los recursos

Las diatomeas viven en ambientes muy predecibles, estables y consistente a lo largo del tiempo

La abundancia total es independiente de qué especies hay

Tienes datos de alta calidad de un ecosistema con patrones ecológicos claros que revelan competencia, estabilidad temporal, y gradientes espaciales interesantes para investigar.

Distribución de fuerza de correlaciones en el Top 5:

Hallazgos

- **TaxonName_num ↔ TaxonCode_num Correlación:** 1.0000 (|1.0000|)

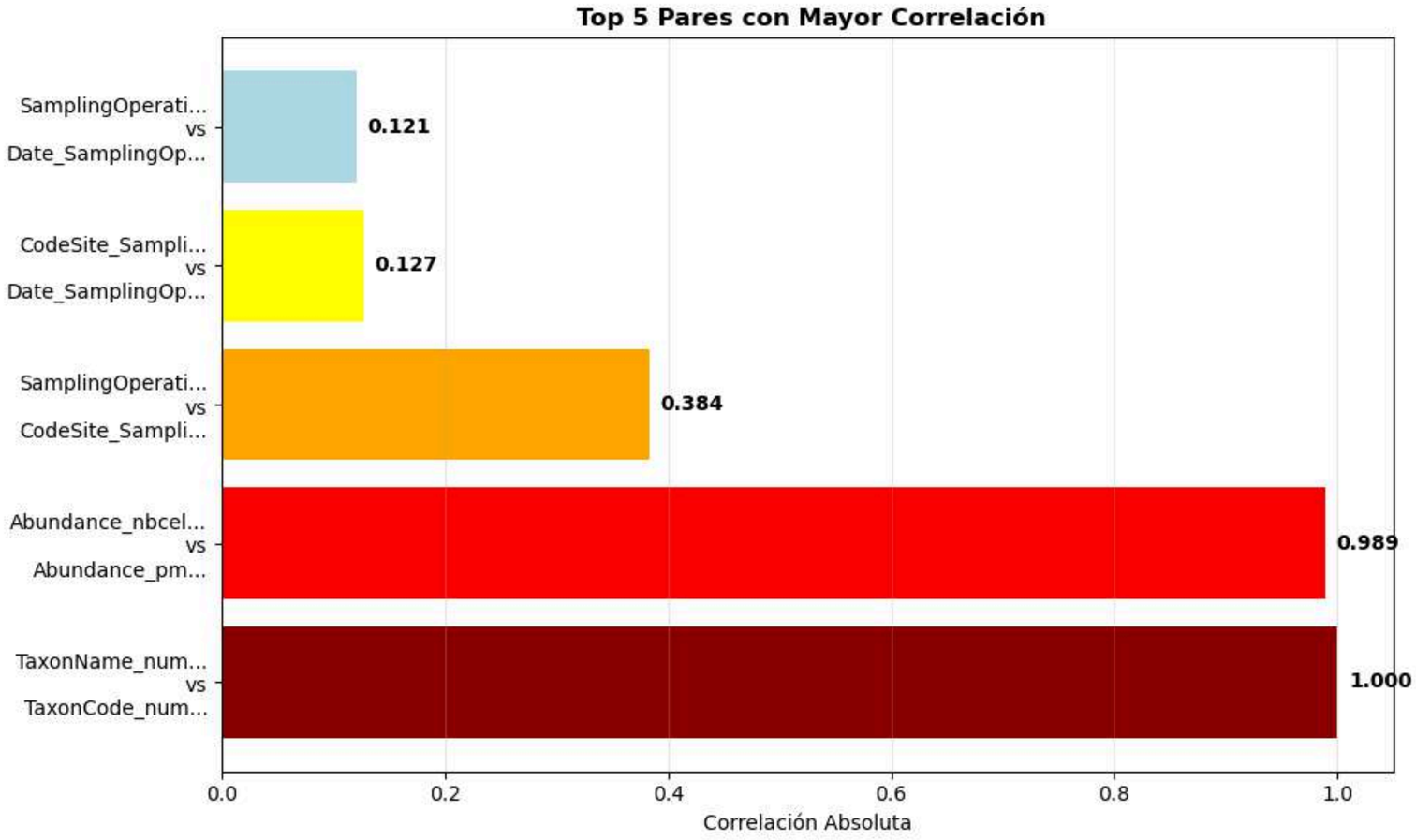
Interpretación: Muy Fuerte Relación POSITIVA: Cuando TaxonName_num aumenta, TaxonCode_num tiende a aumentar 2.

- **Abundance_nbcell ↔ Abundance_pm Correlación:** 0.9890 (|0.9890|)

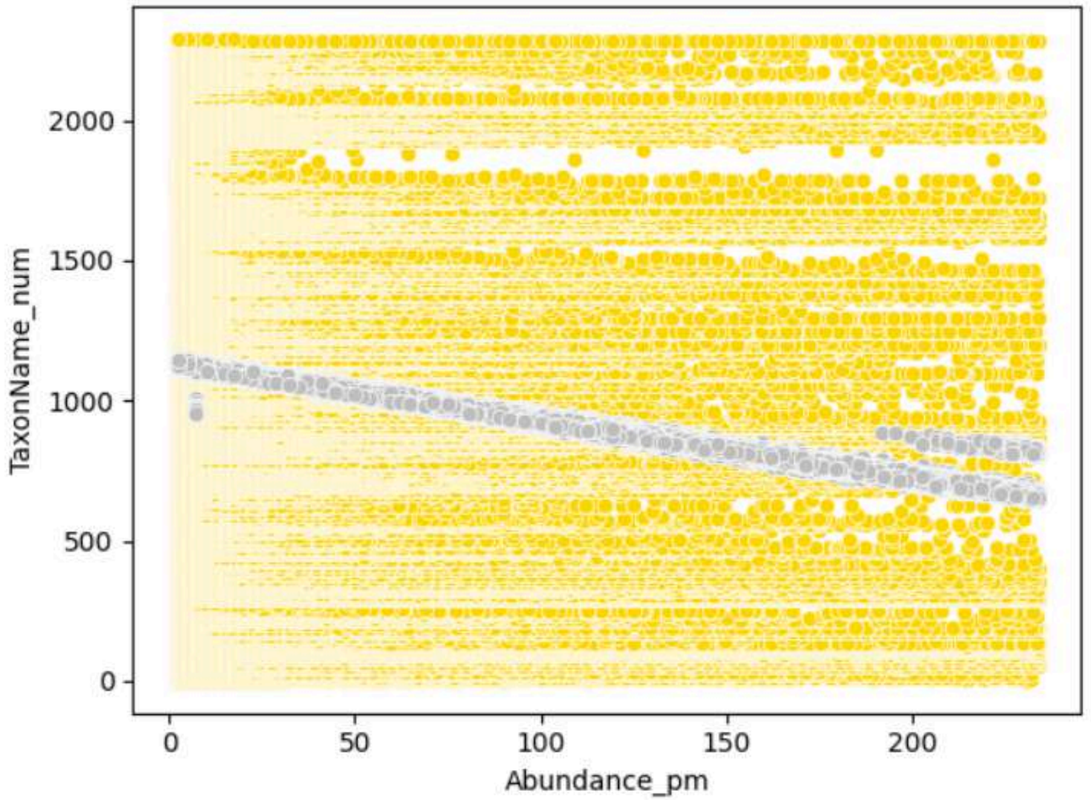
Interpretación: Muy Fuerte Relación POSITIVA: Cuando Abundance_nbcell aumenta, Abundance_pm tiende a aumentar 3.

- **SamplingOperations_code_num ↔ CodeSite_SamplingOperations_num Correlación:** 0.3836 (|0.3836|)

Interpretación: Débil Relación POSITIVA: Cuando SamplingOperations_code_num aumenta, CodeSite_SamplingOperations_num tiende a aumentar



TaxonName

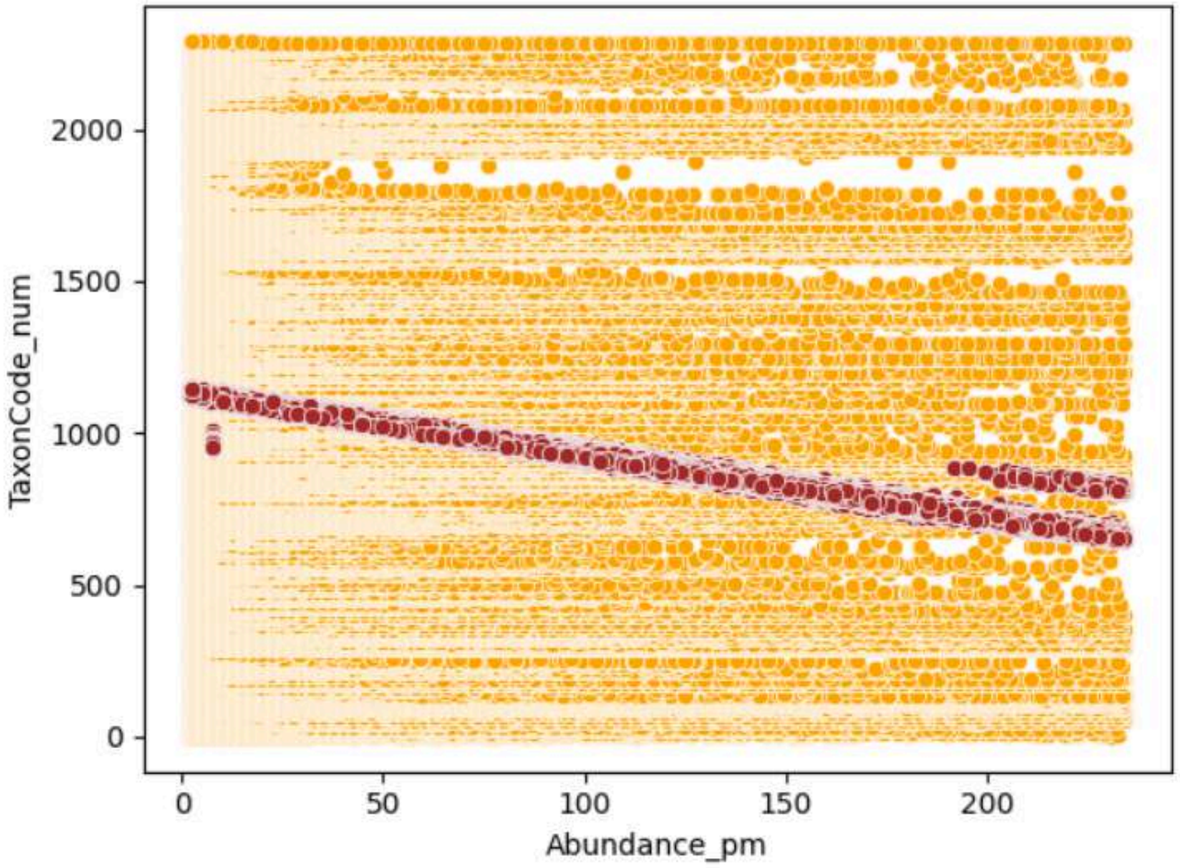


- Se codificaron las variables categóricas en valores numéricos según su frecuencia.
- Se detectaron y ajustaron los valores atípicos.
- Se validó que el dataframe final contuviera únicamente variables numéricas.

Abundance_nbcell	-0.100
Abundance_pm	-0.101
SamplingOperations_code_num	-0.018

Coef de correlación antes	-0.101
Coef de correlación después	0.1016
R^2	0.0103

TaxonCode

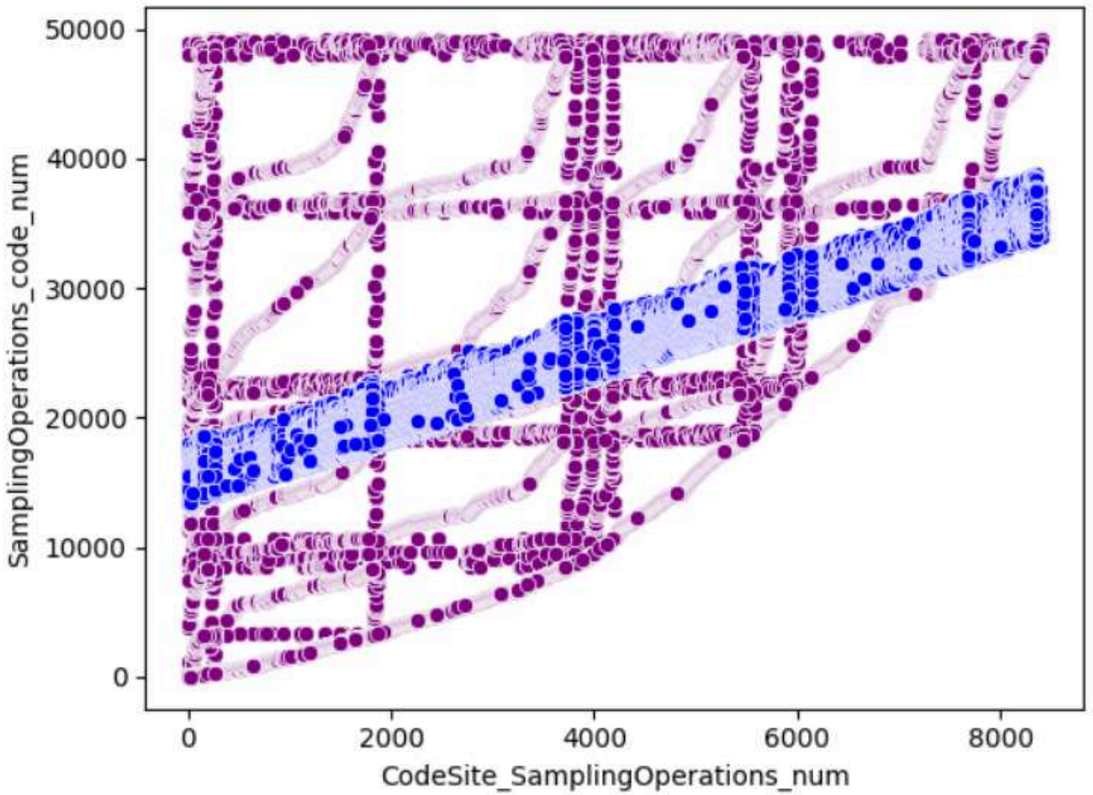


- Se codificaron las variables categóricas en valores numéricos según su frecuencia.
- Se detectaron y ajustaron los valores atípicos.
- Se validó que el dataframe final contuviera únicamente variables numéricas.

Abundance_nbcell	-0.100
Abundance_pm	-0.101
SamplingOperations_code_num	-0.018

Coef de correlación antes	-0.101
Coef de correlación después	0.1016
R^2	0.0103

SamplingOperations_code

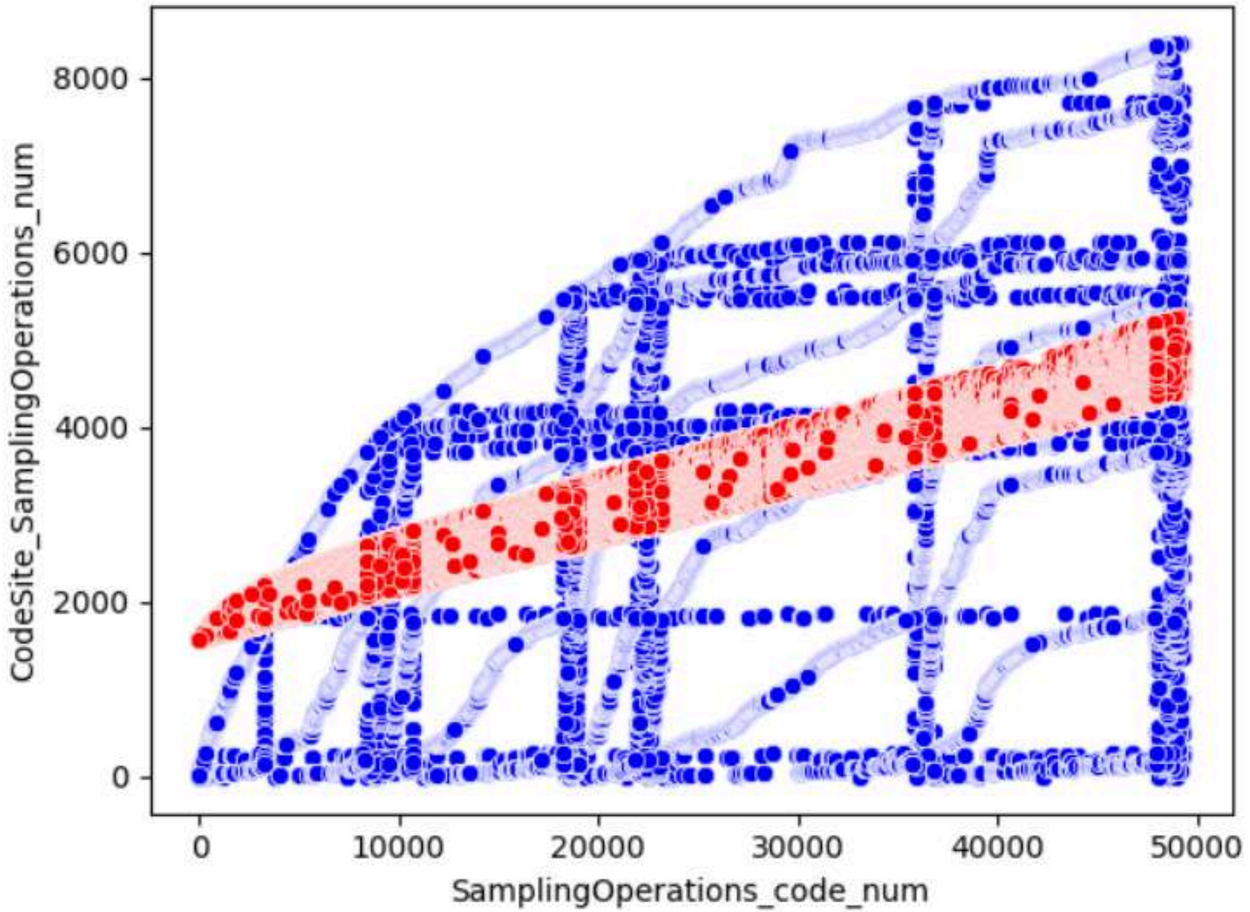


- Se codificaron las variables categóricas en valores numéricos según su frecuencia.
- Se detectaron y ajustaron los valores atípicos.
- Se validó que el dataframe final contuviera únicamente variables numéricas.

CodeSite_SamplingOperations	0.384
Date_SamplingOperation	0.121

Coef de correlación antes	0.3836
Coef de correlación después	0.3903
R^2	0.1524

CodeSite_SamplingOperations

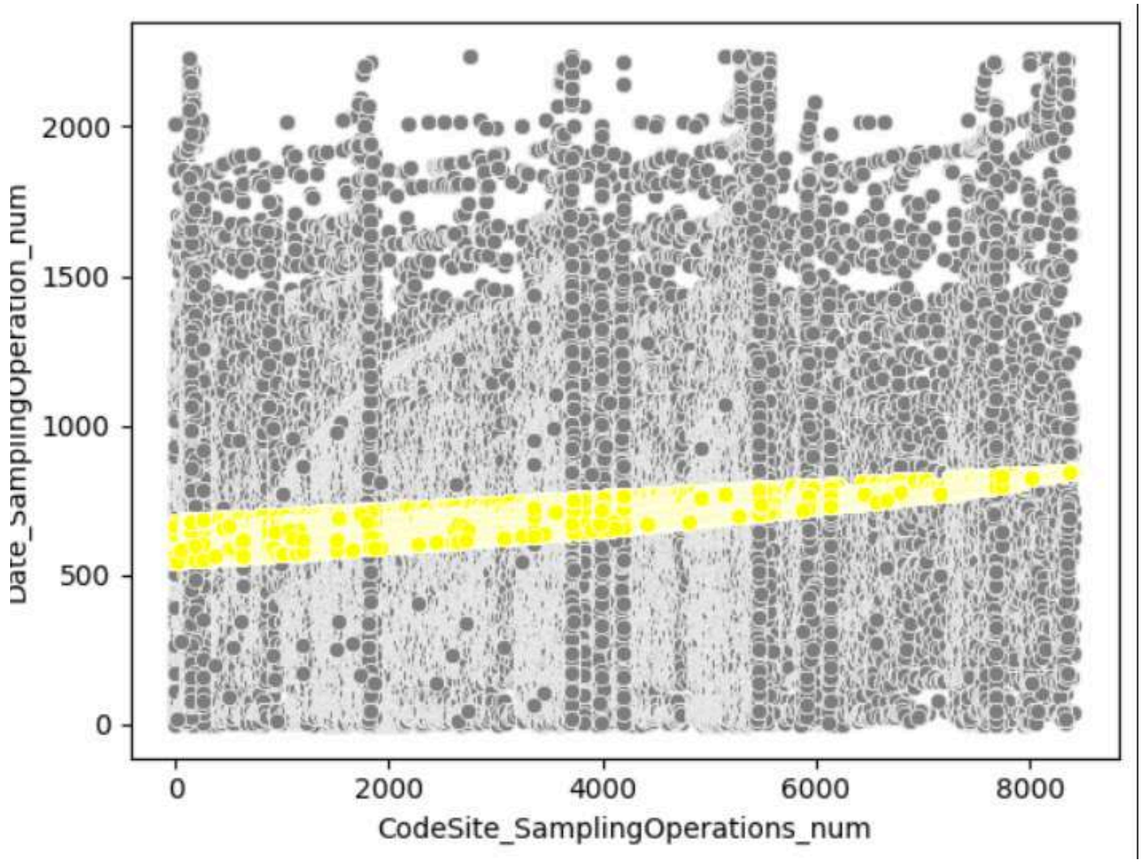


- Se codificaron las variables categóricas en valores numéricos según su frecuencia.
- Se detectaron y ajustaron los valores atípicos.
- Se validó que el dataframe final contuviera únicamente variables numéricas.

SamplingOperations_code	0.384
Date_SamplingOperation	0.127

Coef de correlación antes	0.3836
Coef de correlación después	0.3920
R^2	0.1537

Date_SamplingOperation

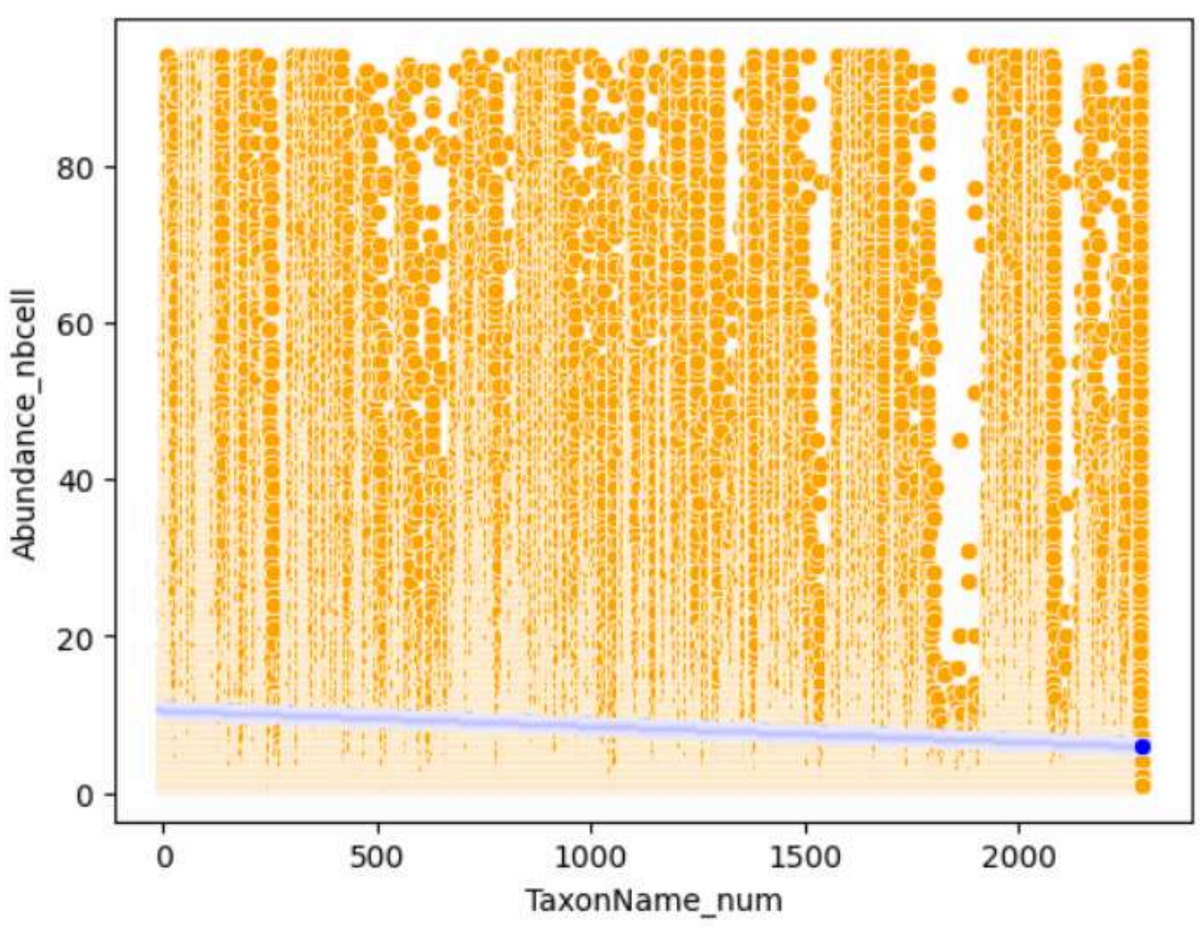


- Se codificaron las variables categóricas en valores numéricos según su frecuencia.
- Se detectaron y ajustaron los valores atípicos.
- Se validó que el dataframe final contuviera únicamente variables numéricas.

SamplingOperations_code	0.121
CodeSite_SamplingOperations	0.127

Coef de correlación antes	0.127
Coef de correlación después	0.149
R^2	0.0222

Abundance_nbcell

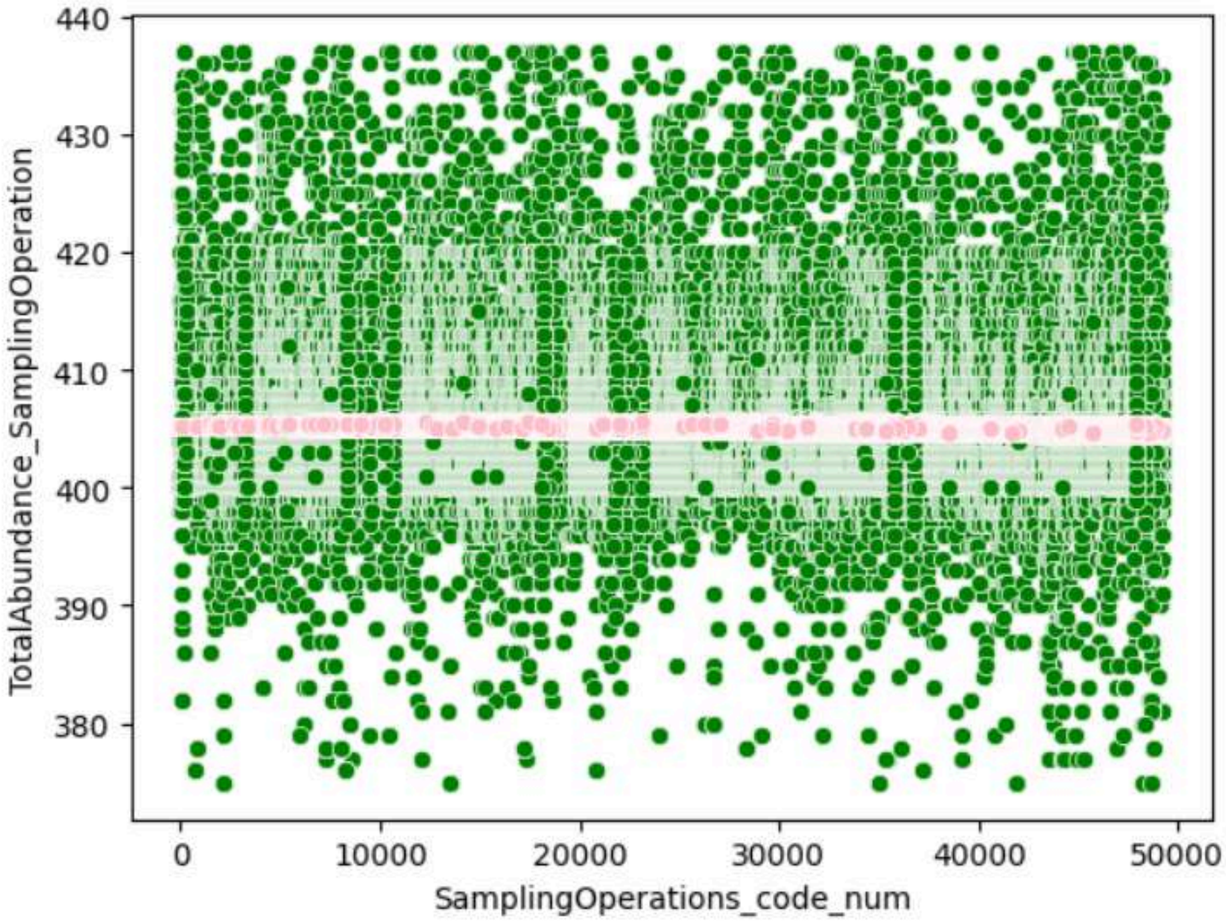


- Se codificaron las variables categóricas en valores numéricos según su frecuencia.
- Se detectaron y ajustaron los valores atípicos.
- Se validó que el dataframe final contuviera únicamente variables numéricas.

TaxonName	-0.100
TaxonCode	-0.100

Coef de correlación antes	-0.100
Coef de correlación después	0.104
R^2	0.0109

TotalAbundance_SamplingOperation

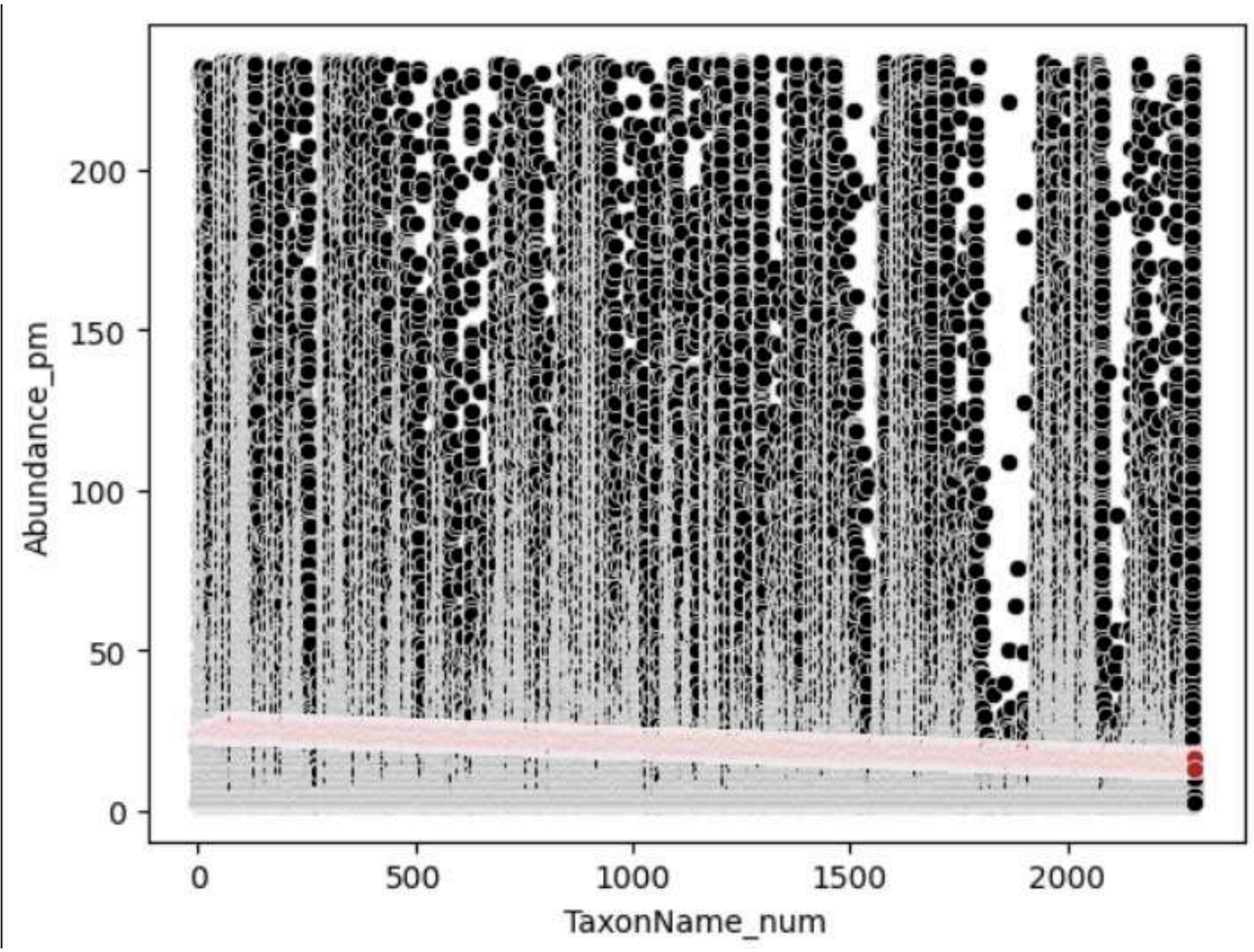


- Se codificaron las variables categóricas en valores numéricos según su frecuencia.
- Se detectaron y ajustaron los valores atípicos.
- Se validó que el dataframe final contuviera únicamente variables numéricas.

Date_SamplingOperation_num	0.008
CodeSite_SamplingOperations	0.015
SamplingOperations_code_num	-0.018

Coef de correlación antes	-0.18
Coef de correlación después	0.029
R^2	0.0008

Abundance_pm



- Se codificaron las variables categóricas en valores numéricos según su frecuencia.
- Se detectaron y ajustaron los valores atípicos.
- Se validó que el dataframe final contuviera únicamente variables numéricas.

TaxonName	-0.101
TaxonCode	-0.101
SamplingOperations_code_num	0.40

Coef de correlación antes	-0.101
Coef de correlación después	0.108
R^2	0.0116

ACTIVIDAD 2.2

FORVIA
faurecia

Transformacion de variables

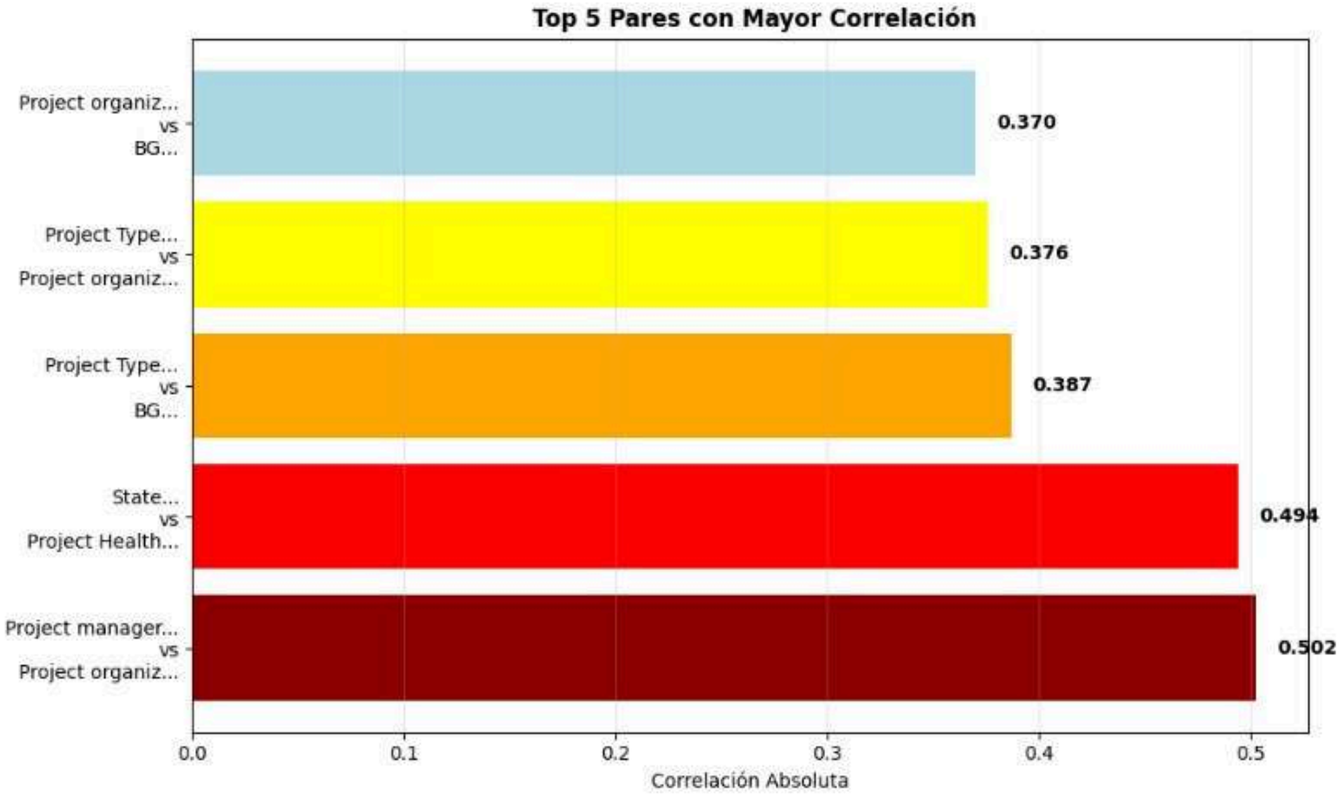
Mapeo con un ciclo for

Index	Project Type	Geographical scope	Project manager	State	Project size	Project Org	BG	Project Health	On-Hold
0	1	63	2	1	3	1	1	1	1
1	1	62	15	1	2	1	2	2	2
2	1	51	20	1	1	1	2	2	1
3	1	51	15	1	3	1	2	1	2
4	1	61	2	1	1	1	2	1	1
...	...	...	...	...	...	...	...	...	...
241	6	43	119	1	2	9	3	1	1
242	8	126	27	1	1	4	3	1	1
243	8	42	27	1	1	4	3	1	1
244	1	42	120	1	3	4	3	1	1
245	12	127	121	4	4	35	11	3	3

Distribución de fuerza de correlaciones en el Top 5:

Hallazgos

- Project manager ↔ Project organization → 0.502 (moderada positiva)
- State ↔ Project Health → 0.494 (moderada positiva)
- Project Type ↔ BG → 0.387 (débil positiva)
- Project Type ↔ Project organization → 0.376 (débil positiva)
- Project organization ↔ BG → 0.370 (débil positiva)

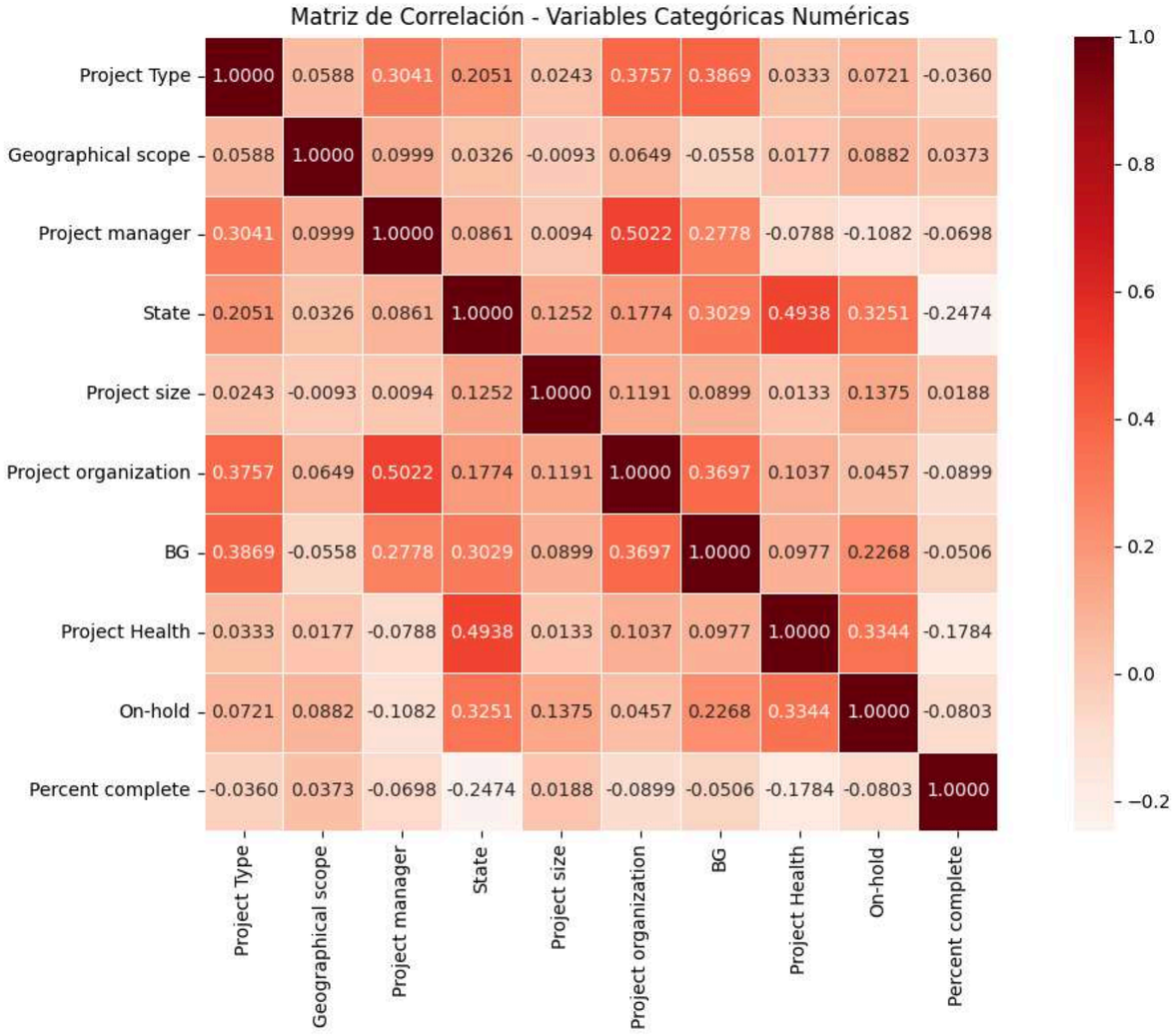


ANÁLISIS DE CORRELACIONES USANDO HEATMAP

Hallazgos de FORVIA

Interpretación de la Matriz de Correlación

- Relaciones moderadas:
- Project manager ↔ Project organization (0.502): A mayor número de Project managers, más proyectos por organización.
- State ↔ Project Health (0.494): Proyectos en estados avanzados tienden a tener mejor salud.
- Relaciones débiles:
- Project Type ↔ BG (0.387), Project Type ↔ Project organization (0.376), Project organization ↔ BG (0.370): Leve asociación entre tipo de proyecto, organización y unidad de negocio.
- Correlaciones muy bajas:
- Percent complete tiene muy poca relación con otras variables, lo que sugiere que el avance de los proyectos no está determinado por estas variables.



ANÁLISIS DE INSIGHTS

Patrones de Registro

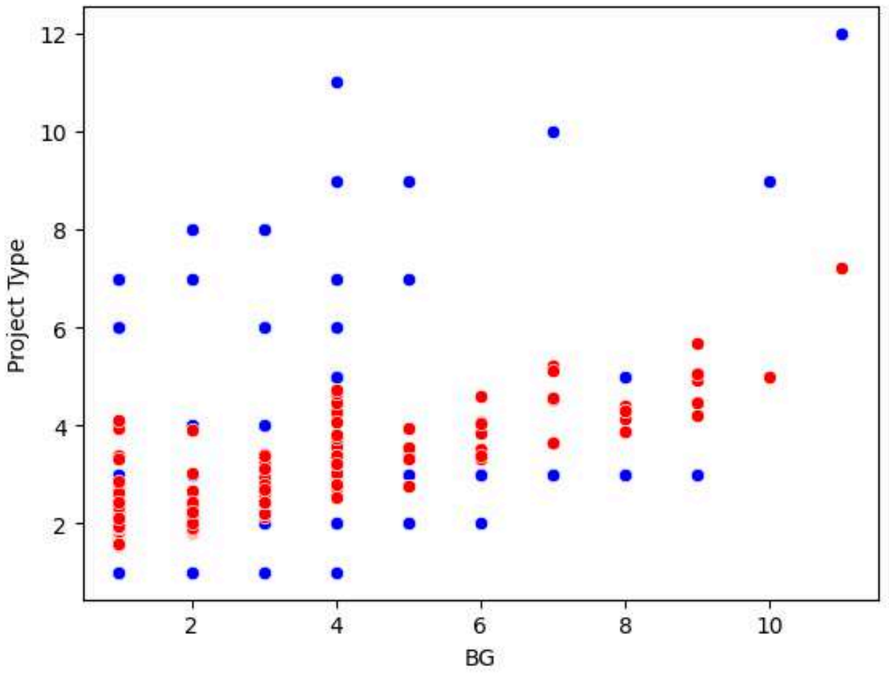
Estandarización de Procesos

Campos No Independientes

Correlación #1: Project Manager ↔ Project Organization (0.505)

Forvia es una organización madura con sistemas sofisticados de gestión de proyectos, operaciones globales diversificadas, y una estructura organizacional adaptable que les permite manejar eficientemente un portfolio complejo y variado de proyectos.

Project Type ↔ BG

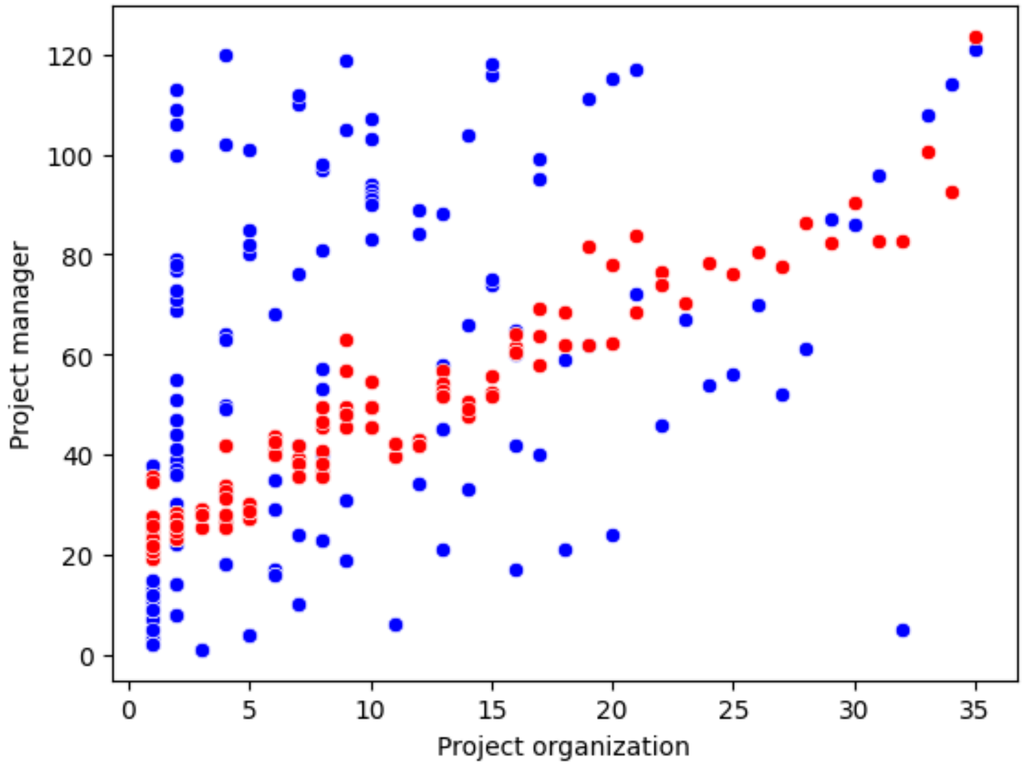


- Se observó una relación positiva débil entre BG y Project Type.
- Los datos reales presentan alta dispersión, sobre todo en valores bajos de BG.
- El modelo múltiple logra ajustar una tendencia ascendente, aunque limitada por la variabilidad.

Project manager	0.3041
Project organization	0.3757
BG	0.3869

Coef de correlación antes	0.3869
Coef de correlación después	0.4645
R^2	0.2157

Project manager ↔ Project organization

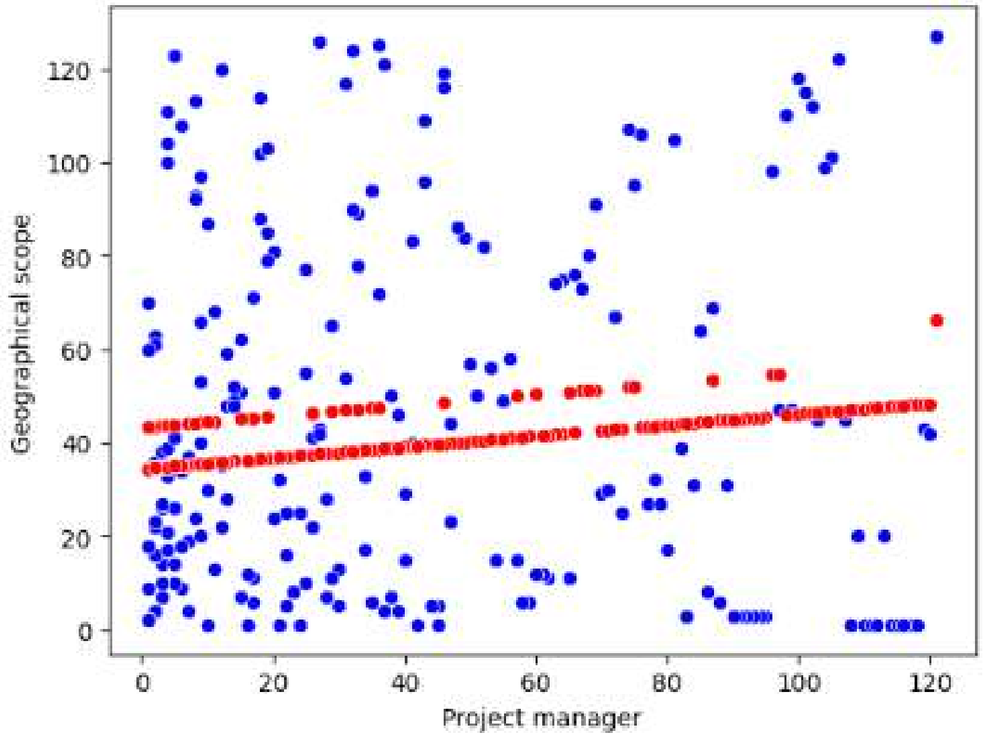


- Se observó una relación positiva moderada entre Project organization y Project manager.
- Los valores reales muestran dispersión alta, aunque la tendencia ascendente es clara.
-
- El modelo múltiple refuerza la relación, pero la variabilidad limita la precisión.

Project Type	0.3041
Project organization	0.5022
BG	0.2778

Coef de correlación antes	0.5022
Coef de correlación después	0.5216
R^2	0.2720

Geographical scope ↔ Project manager

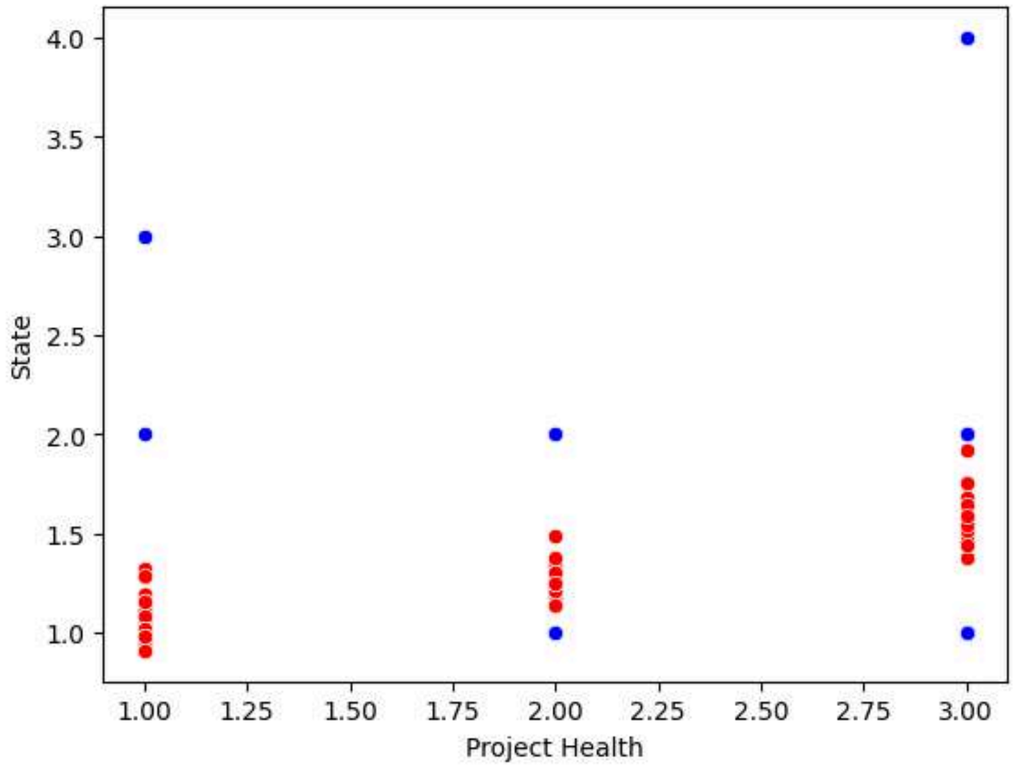


- Se identificó una relación positiva débil entre Project manager y Geographical scope.
- Los valores reales presentan alta dispersión, sin un patrón lineal claro.
- El modelo múltiple ajusta una tendencia ascendente ligera, pero con bajo poder predictivo.

Project manager	0.0999
On-hold	0.0882

Coef de correlación antes	0.0999
Coef de correlación después	0.1410
R^2	0.0198

State ↔ Project Health

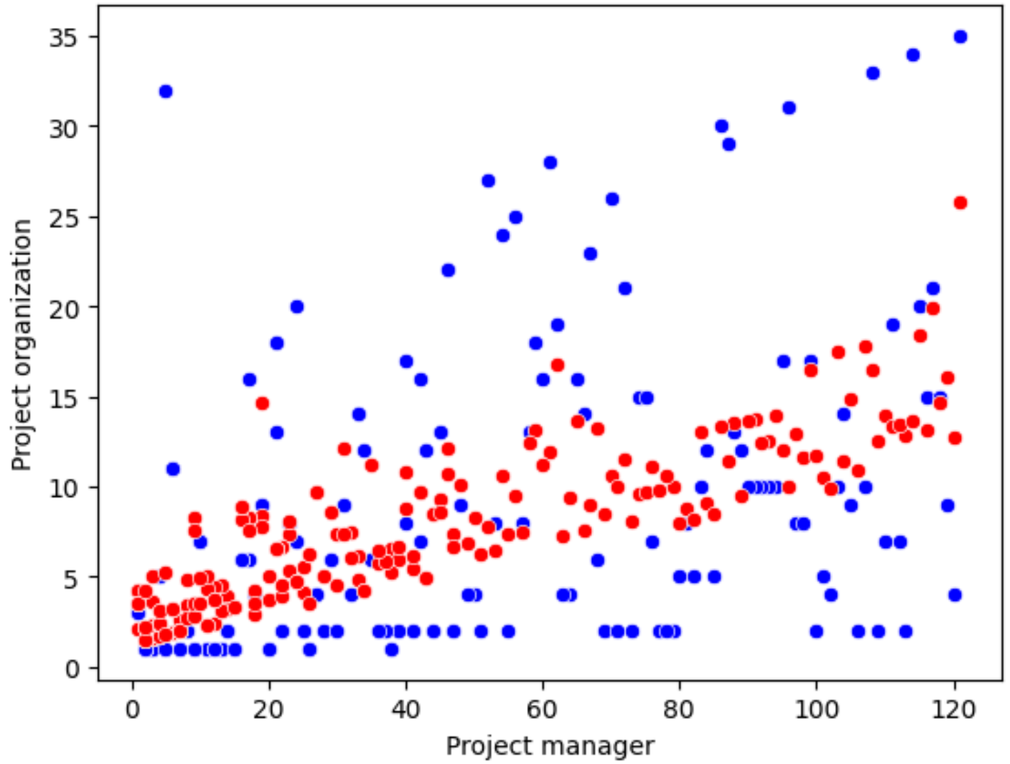


- Se identificó una relación positiva moderada entre State y Project Health.
- La mayoría de proyectos se concentran en estados bajos, con distintos niveles de salud.
- El modelo múltiple confirma la tendencia ascendente, aunque la dispersión reduce su precisión.

On-hold	0.3251
Project Health	0.4938
BG	0.3029

Coef de correlación antes	0.4938
Coef de correlación después	0.5688
R^2	0.3256

Project Organization↔ Project Manager

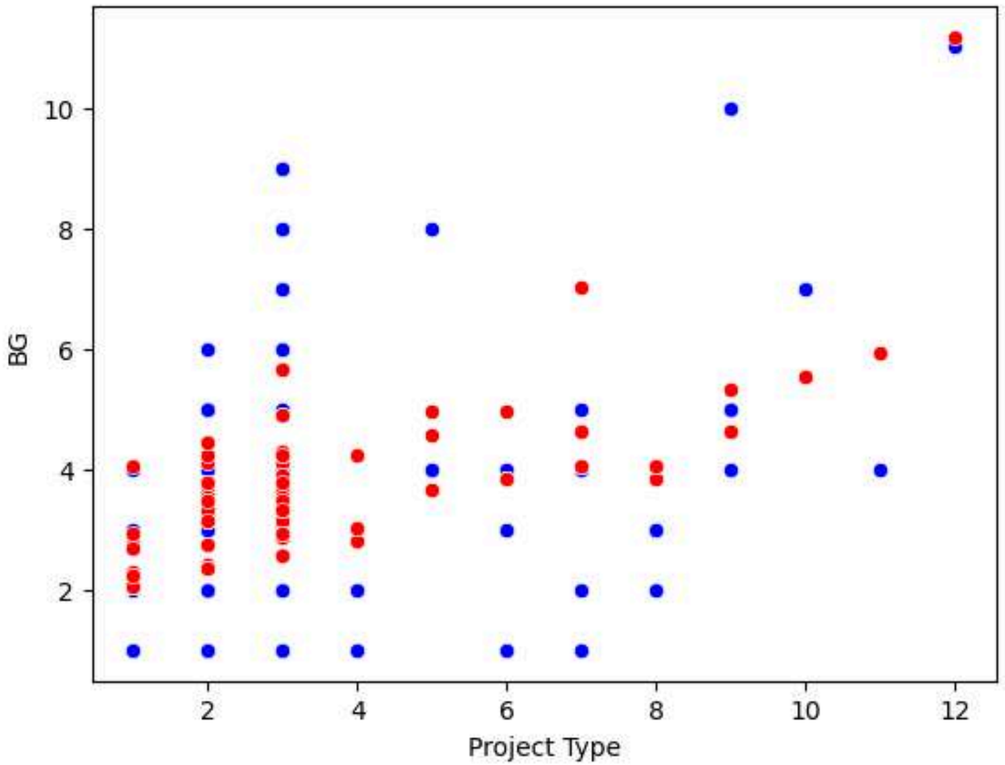


- Se observó una relación positiva moderada entre Project Organization y Project Manager.
- Los datos reales presentan alta dispersión, especialmente en proyectos con valores bajos de Project Organization.
- El modelo múltiple muestra una tendencia ascendente moderada, pero con bajo poder explicativo

Project type	0.3757
Project manager	0.5022
BG	0.3697

Coef de correlación antes	0.5022
Coef de correlación después	0.5801
R^2	0.3365

BG ↔ Project Type

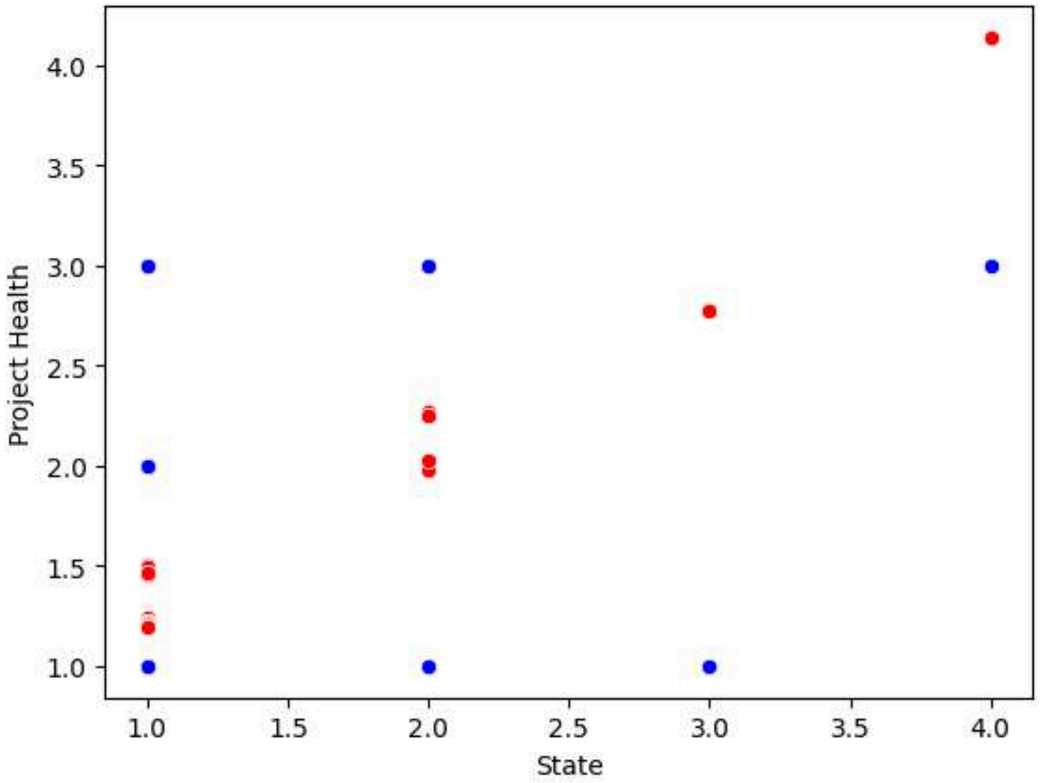


- Se observó una relación positiva débil entre Project Type y BG.
- Los datos reales presentan alta dispersión, sobre todo en valores bajos de Project Type.
- El modelo múltiple muestra una tendencia ascendente ligera, pero con bajo poder explicativo.

Project type	0.3869
Project organization	0.3697
State	0.3029

Coef de correlación antes	0.3869
Coef de correlación después	0.4994
R^2	0.2494

Project Health ↔ State

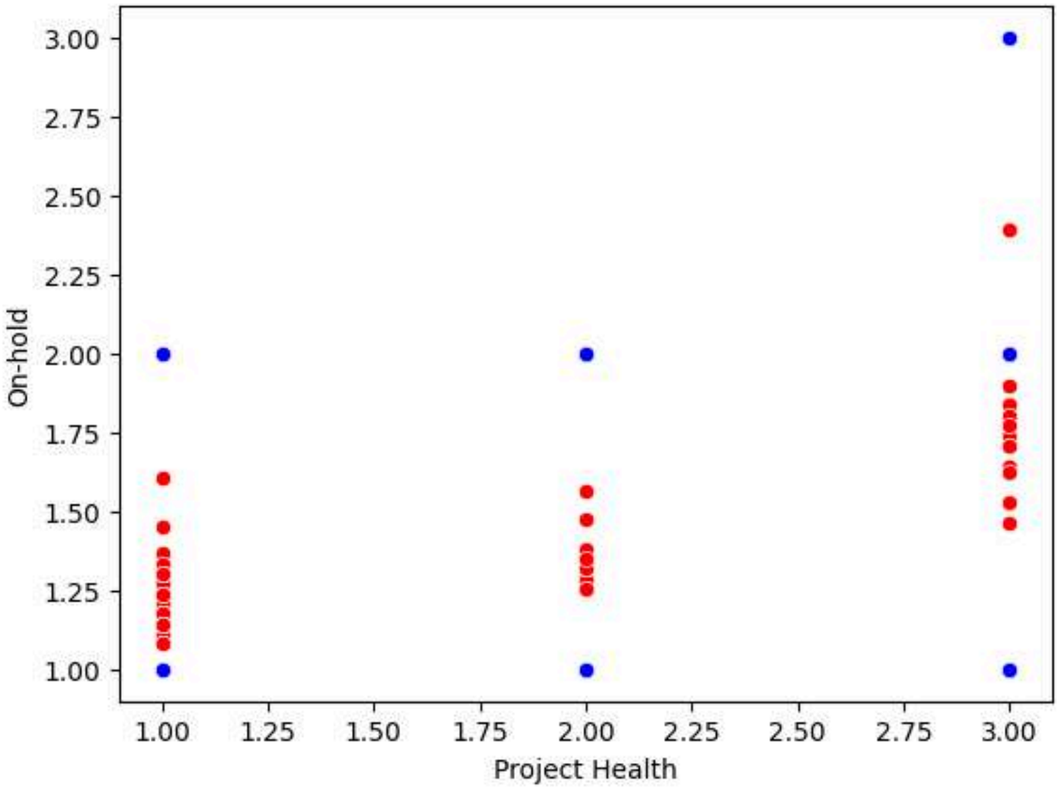


State	0.4938
On-hold	0.3344

- Se encontró una relación positiva moderada entre State y Project Health.
- La mayoría de proyectos se concentran en estados bajos, con diferentes niveles de salud.
- El modelo múltiple confirma la tendencia ascendente, aunque la dispersión limita la precisión.

Coef de correlación antes	0.4938
Coef de correlación después	0.5272
R^2	0.2779

On-hold ↔ Project Health

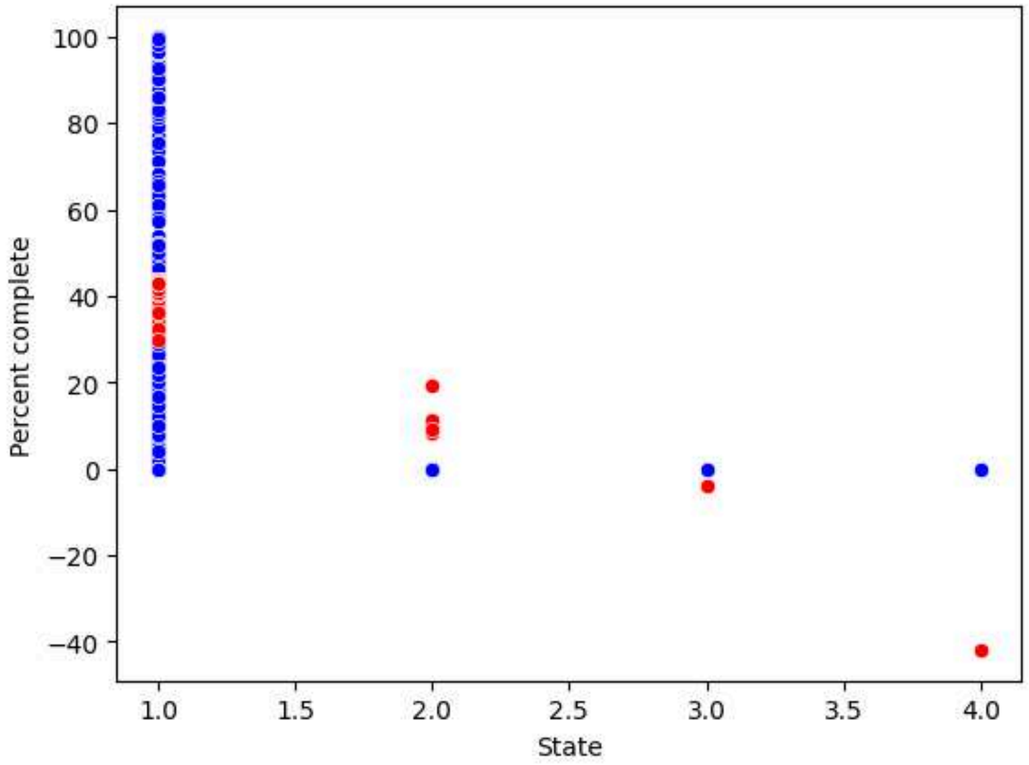


- Se observó una relación positiva débil entre On-hold y Project Health.
- La mayoría de proyectos se concentran en niveles bajos de On-hold.
- Algunos proyectos con buena salud aparecen en pausa, pero el patrón general es poco significativo.

State	0.3251
Project Health	0.3344
BG	0.2268

Coef de correlación antes	0.3344
Coef de correlación después	0.4091
R^2	0.1673

Percent complete ↔ State



State	-0.2474
Project organization	-0.0899
Project Health	-0.1784

- Se identificó una relación positiva muy débil entre State y Percent complete.
- La mayoría de proyectos están en State = 1, con avances muy variados (0% a 100%).
- En estados más altos hay pocos proyectos y con avances bajos.
- No existe un patrón lineal claro → baja utilidad predictiva.

Coef de correlación antes	-0.2474
Coef de correlación después	0.2597
R^2	0.0674

¡MUCHAS GRACIAS POR SU ATENCION!