Octubre 2025

Análisis exploratorio de proyectos Datathon y FORVIA

REGRESIÓN LINEAL SIMPLE Y MÚLTIPLE

fORVIANOS.py

FORVIANOS.PY

Maria Matanzo

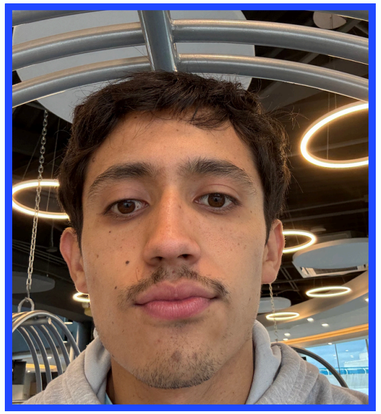
A01737554

Jorge Cortes

A01736236

Marco Cornejo

A01276411



Eduardo Torres

A01734935

Laisha Puan

A01736397

OBJETIVOS

- Aplicar técnicas de regresión lineal simple y múltiple para analizar relaciones entre variables cuantitativas de un dataset real.
- Transformar variables categóricas en numéricas utilizando codificación basada en jerarquía de frecuencias, para preparar los datos para modelos estadísticos.
- Identificar correlaciones significativas entre variables mediante un mapa de calor (heatmap) y seleccionar los pares con mayor relación.
- Construir y evaluar modelos de regresión múltiple, interpretando los coeficientes obtenidos y su relación con los resultados visualizados en los mapas de calor.

METODOLOGÍA

Preparación de los datos

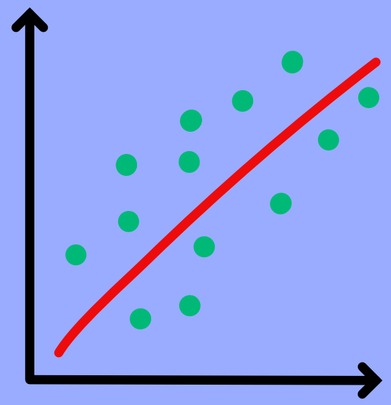
01 Uso del dataset 01_DiatomInventories_GTstudentproject_B.csv.

02 Conversión de variables categóricas: TaxonName, TaxonCode, SamplingOperations_code, CodeSite_SamplingOperations, Date_SamplingOperation.

03 Obtención de un DataFrame solo con variables numéricas.

METODOLOGÍA

Regresión Lineal Simple



01 Cálculo de correlaciones.

02 Selección de los 5 pares de variables con mayor correlación.

03 Visualización mediante heatmap.

METODOLOGÍA

Regresión Lineal Múltiple

01 Construcción de modelos predictivos para variables cuantitativas clave (Abundance_nbcell, TotalAbundance_SamplingOperation, Abundance_pm).

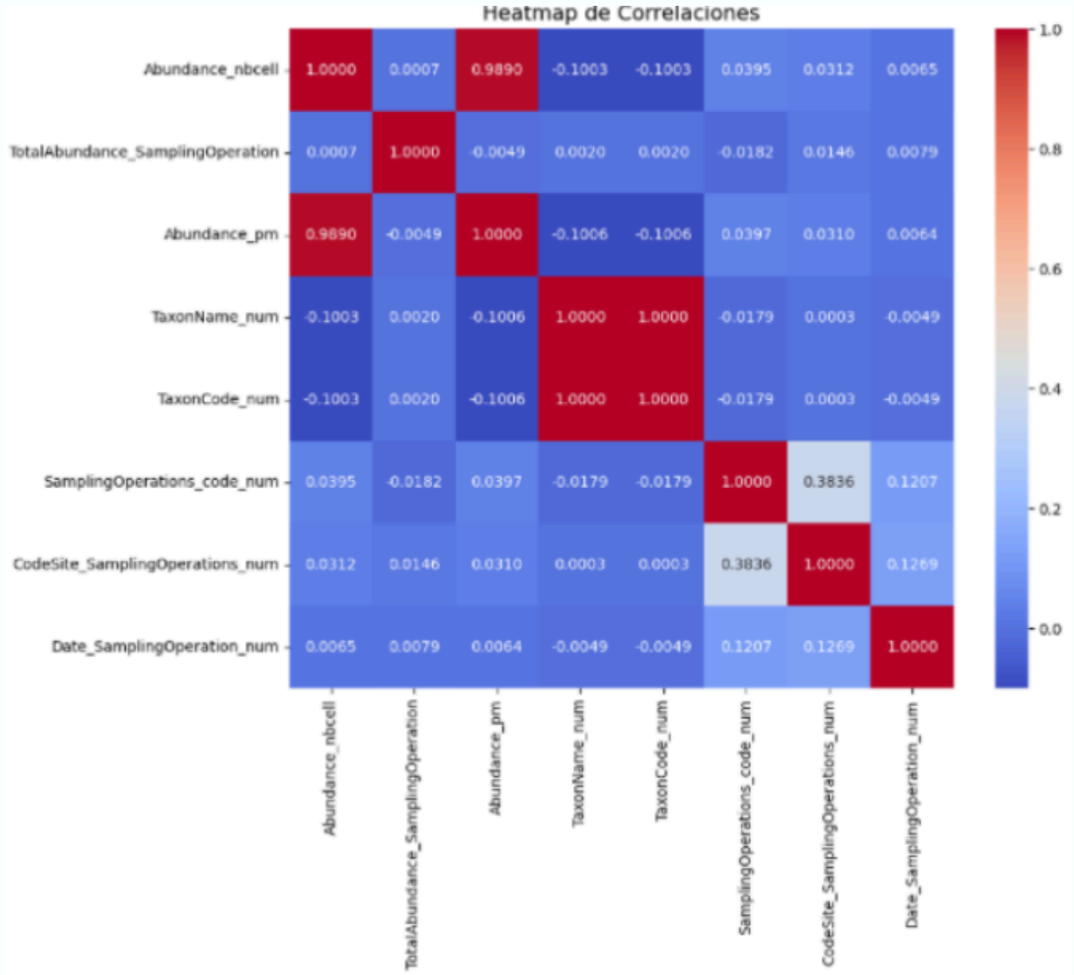
02 Comparación de coeficientes y validación de los modelos.

ACTIVIDAD 2_1

fORVIANOS.py

MODELO PREDICTIVO

- **Gráfica:** Heatmap de correlaciones
- **Hallazgos:**
  - Se identificaron variables con alta correlación, como Abundance_nbcell y Abundance_pm.
  - TaxonName_num y TaxonCode_num mostraron una correlación perfecta (1.0), indicando redundancia.
  - La mayoría de las demás variables tienen correlaciones bajas, lo que limita el poder explicativo de los modelos lineales.
  - Las variables seleccionadas para la regresión múltiple se eligieron en función de estas correlaciones.



MODELO 1

Variables independientes (X):

- Abundance_pm
- Abundance_nbcell
- SamplingOperations_code_num

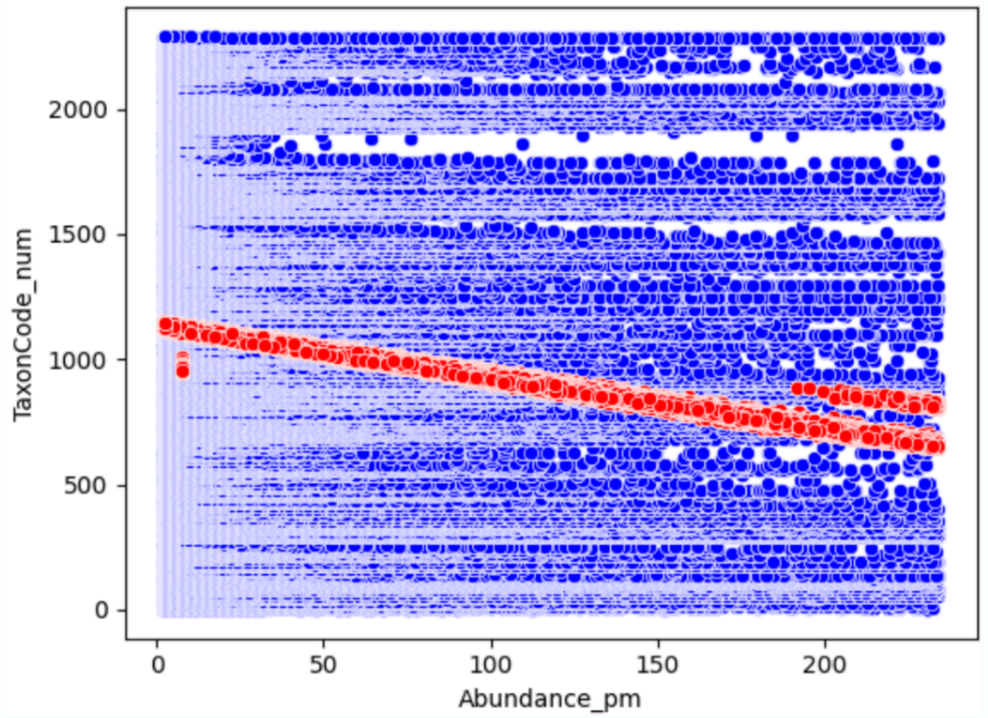
Variable dependiente (Y):

- TaxonName_num (codificada numéricamente).

MODELO 1

- **Gráfica:** Dispersión de valores reales vs. predicciones

- **Hallazgos:**
  - Los puntos azules representan los valores reales de TaxonName_num en función de Abundance_pm.
  - Los puntos rojos corresponden a las predicciones del modelo (Predicciones1).
  - El modelo lineal muestra un ajuste pobre: la recta de predicción no sigue el patrón real de los datos.
  - Esto confirma que la regresión lineal no es adecuada para explicar la variabilidad en esta variable.



MODELO 2

Variables independientes (X):

- Abundance_pm
- Date_SamplingOperation_num
- CodeSite_SamplingOperations_num

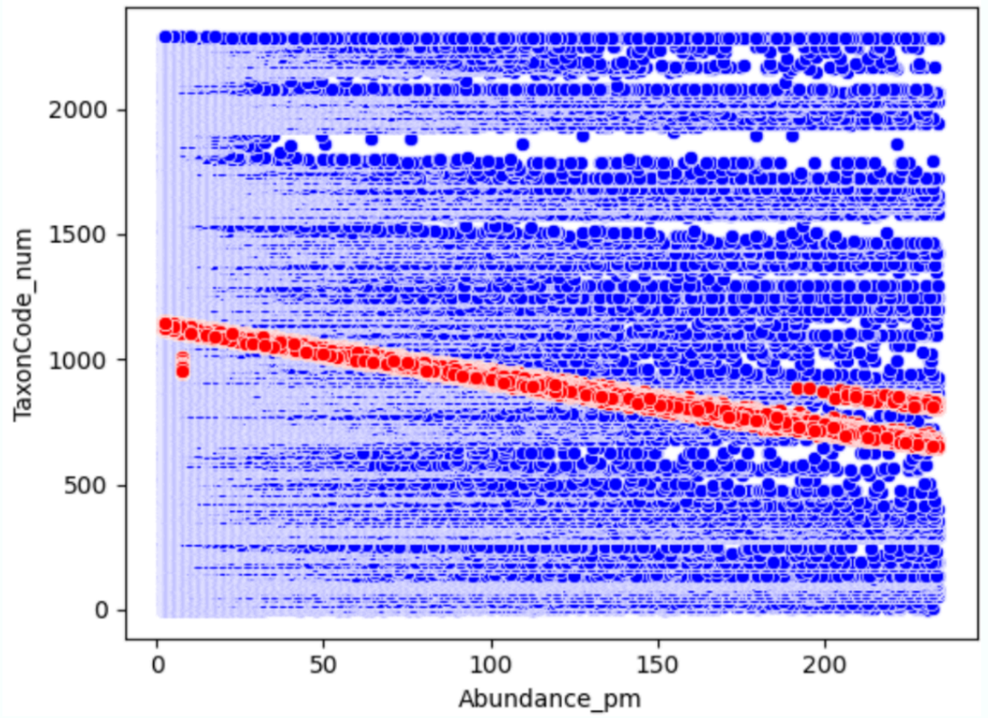
Variable dependiente (Y):

- SamplingOperations_code_num

MODELO 2

- **Gráfica:** Dispersión de valores reales vs. predicciones

- **Hallazgos:**
  - Los puntos azules representan los valores reales de TaxonName_num en función de Abundance_pm.
  - Los puntos rojos corresponden a las predicciones del modelo (Predicciones1).
  - El modelo lineal muestra un ajuste pobre: la recta de predicción no sigue el patrón real de los datos.
  - Esto confirma que la regresión lineal no es adecuada para explicar la variabilidad en esta variable.



MODELO 3

Variables independientes (X):

- SamplingOperations_code_num: código de la operación de muestreo.
- Date_SamplingOperation_num: fecha de muestreo en formato numérico.
- Abundance_nbcell: número de células (abundancia).

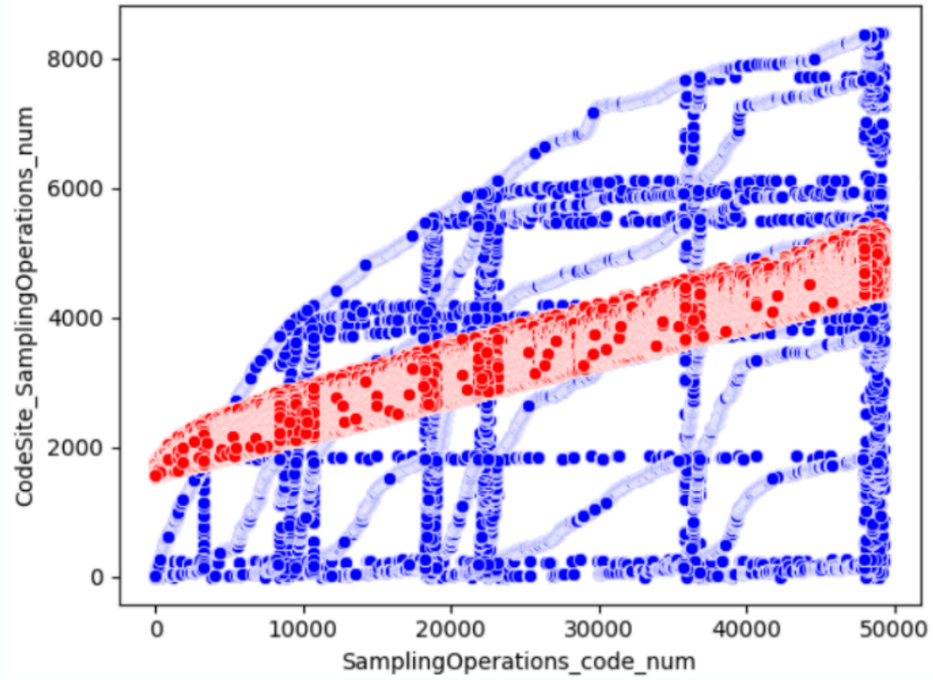
Variable dependiente (Y):

- CodeSite_SamplingOperations_num: código del sitio de muestreo.

MODELO 3

- **Gráfica:** Dispersión de valores reales vs. predicciones

- **Hallazgos:**
  - Los puntos azules representan los valores reales de CodeSite_SamplingOperations_num en función de SamplingOperations_code_num.
  - Los puntos rojos corresponden a las predicciones del modelo (Predicciones3).
  - El patrón real de los datos es escalonado y disperso, con múltiples rangos y agrupaciones.
  - Las predicciones forman una franja recta que no sigue la complejidad de los valores reales.
  - Esto confirma que la regresión lineal simplifica demasiado el fenómeno, generando resultados de baja utilidad predictiva.



MODELO 4

Variables independientes (X):

- CodeSite_SamplingOperations_num
- SamplingOperations_code_num
- Abundance_pm

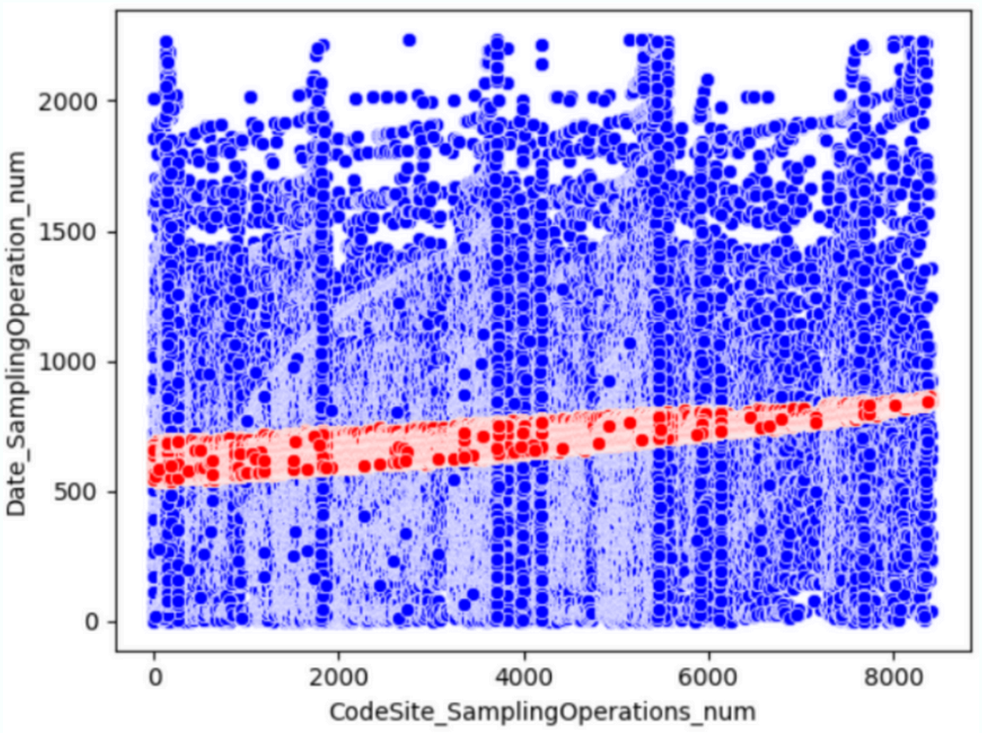
Variable dependiente (Y):

- Date_SamplingOperation_num

MODELO 4

- **Gráfica:** Dispersión

- **Hallazgos:**
  - Los puntos azules son muy dispersos y no muestran un patrón definido.
  - Las fechas varían según el calendario, no por la abundancia o los códigos de muestreo.
  - El modelo (puntos rojos) dibuja una recta que no logra capturar la variación real de las fechas.

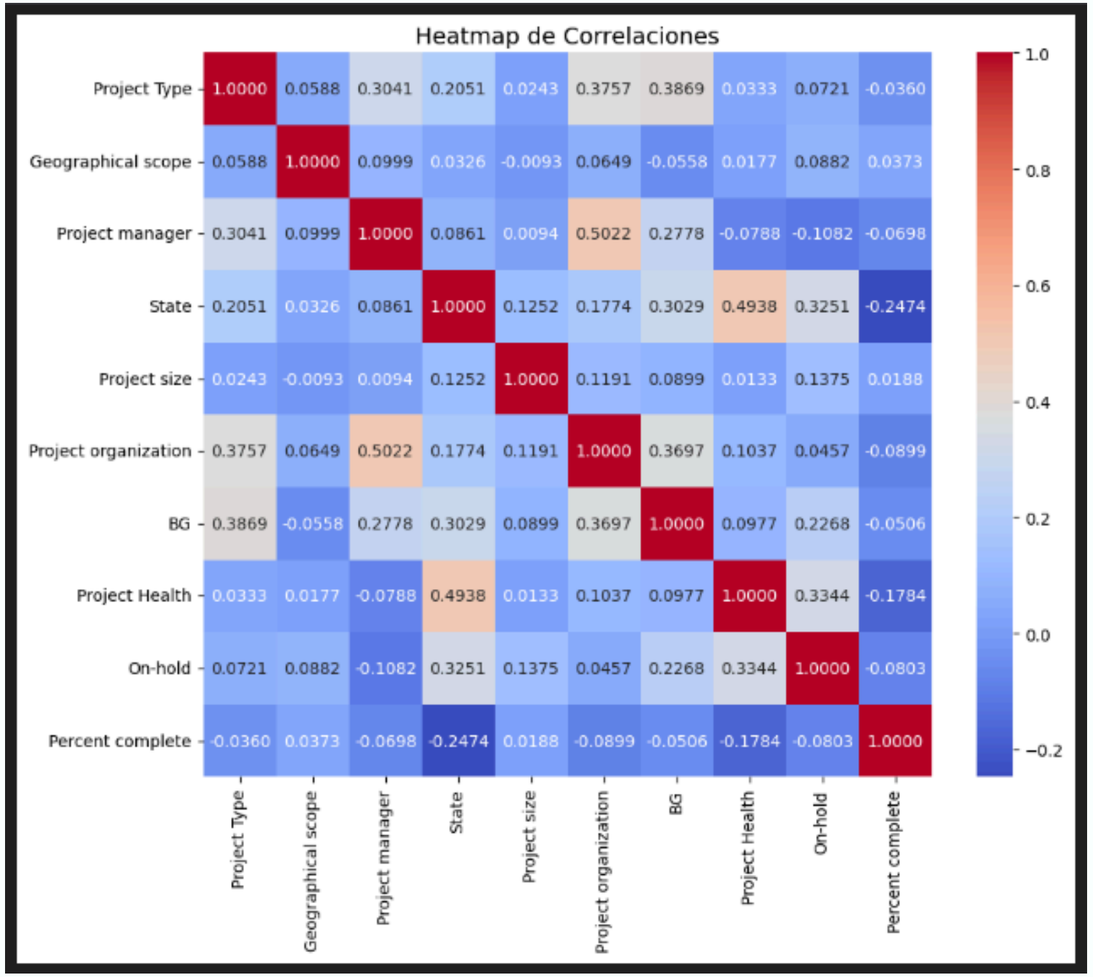


ACTIVIDAD 2_2

fORVIANOS.py

MODELO PREDICTIVO

- **Gráfica:** Heatmap de correlaciones
- **Hallazgos:**
  - Se identificaron variables con correlaciones destacadas, como Project manager y Project organization (0.50), y State con Project Health (0.49).
  - También se observó relación entre Project Type y BG (0.38), mientras que State y Percent complete mostraron correlación negativa (-0.25).
  - La mayoría de las demás variables presentan correlaciones bajas, por lo que su influencia es limitada.
  - Las variables más correlacionadas servirán como base para el análisis y los modelos de regresión múltiple.



MODELO 1

Variables independientes (X):

- Project manager: responsable de la gestión del proyecto.
- On-hold: indicador de si el proyecto está en pausa o detenido.
- Project organization: organización del proyecto.

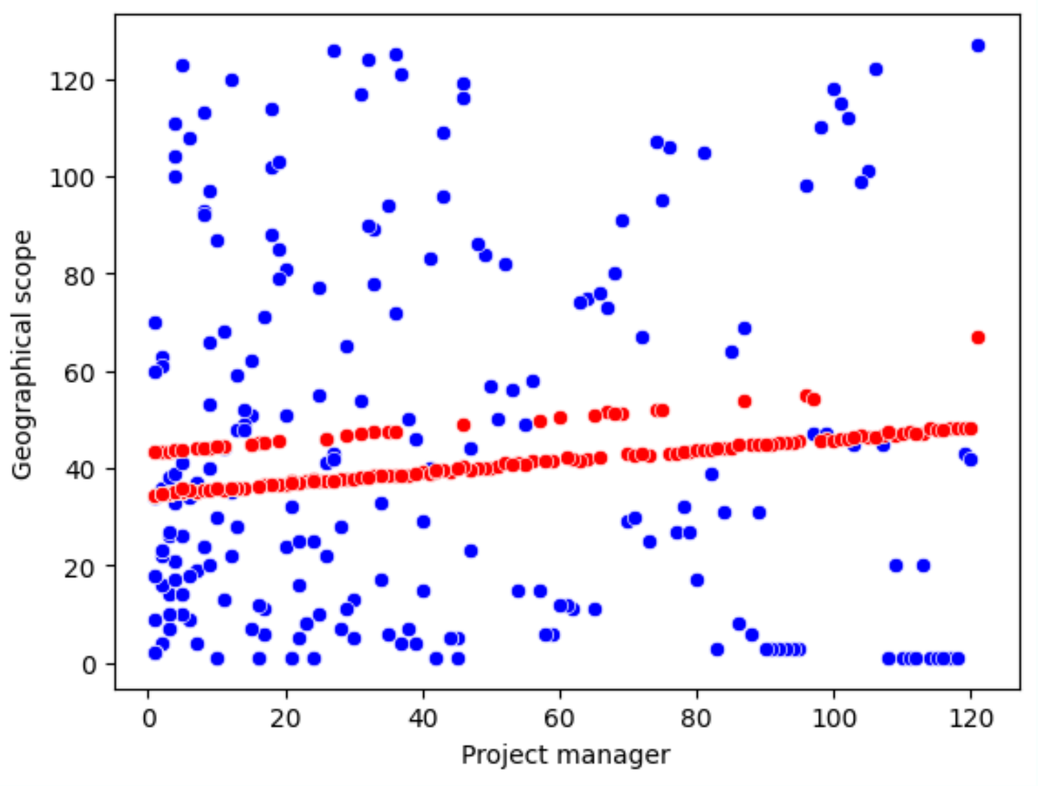
Variable dependiente (Y):

- Geographical scope: alcance geográfico del proyecto.

MODELO 1

- **Gráfica:** Dispersión

- **Hallazgos:**
  - Los puntos azules representan los valores reales de CodeSite_SamplingOperations_num en función de SamplingOperations_code_num.
  - Los puntos rojos corresponden a las predicciones del modelo de regresión lineal.
  - El patrón real de los datos es escalonado y disperso, con múltiples rangos y agrupaciones.
  - Las predicciones forman una franja recta que no sigue la complejidad de los valores reales.
  - Esto confirma que la regresión lineal simplifica demasiado el fenómeno, generando resultados de baja utilidad predictiva.



MODELO 2

Variables independientes (X):

- Project manager: responsable de la gestión del proyecto.
- Project Type: tipo de proyecto.
- BG: código de la unidad de negocio (Business Group).

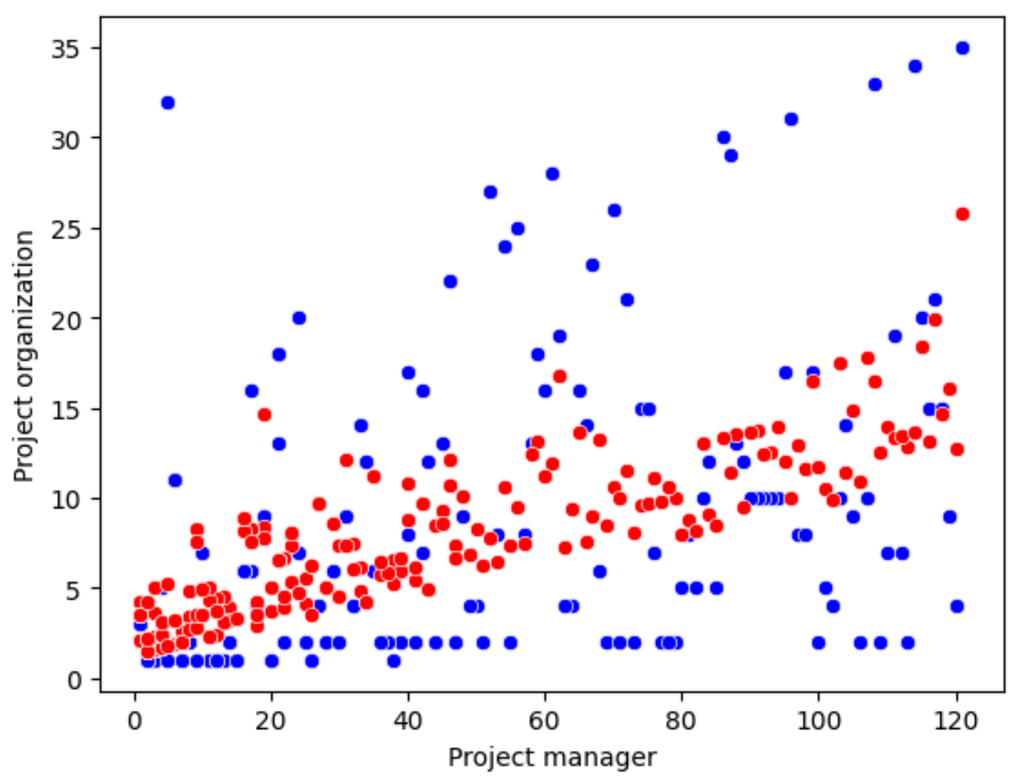
Variable dependiente (Y):

- Project organization: organización del proyecto.

MODELO 2

- **Gráfica:** Dispersión

- **Hallazgos:**
  - Los puntos azules representan los valores reales de Project organization en función de Project manager.
  - Los puntos rojos corresponden a las predicciones realizadas por el modelo.
  - Se observa que los datos reales muestran alta dispersión y variabilidad, con valores que no siguen un patrón lineal definido.
  - Las predicciones forman una franja ascendente más ordenada, pero que no logra capturar del todo la complejidad y dispersión de los datos observados.
  - Esto evidencia que, aunque se detecta una ligera tendencia positiva, el modelo no explica adecuadamente la relación entre las variables y ofrece una capacidad predictiva limitada.



MODELO 3

Variables independientes (X):

- Project organization: organización del proyecto.
- Project Type: tipo de proyecto.
- BG: código de la unidad de negocio (Business Group).

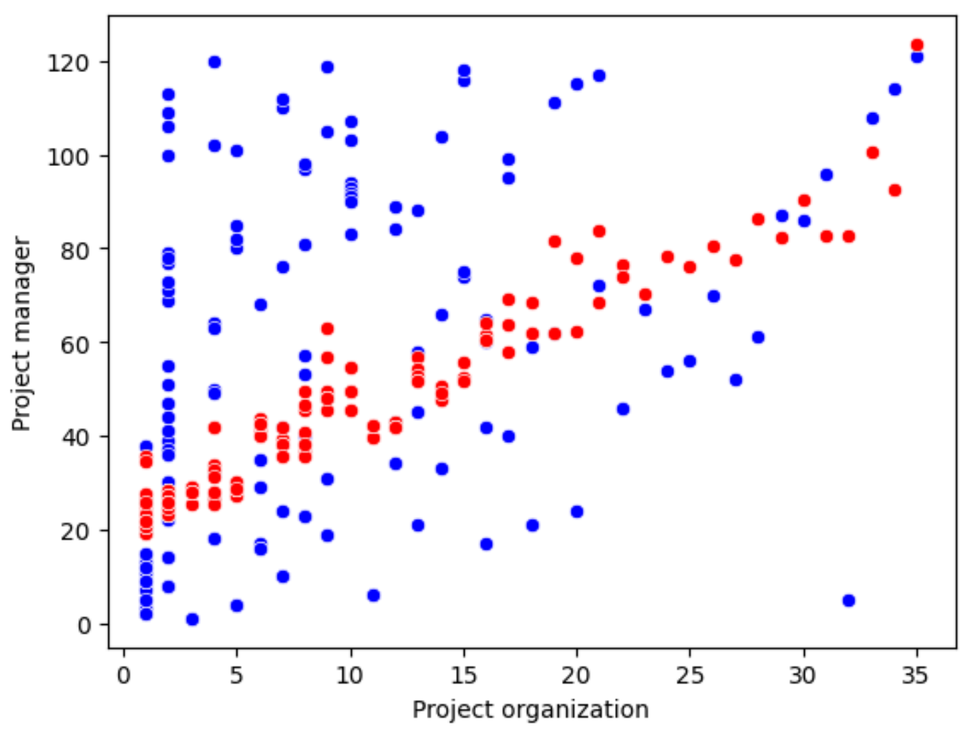
Variable dependiente (Y):

- Project manager: responsable de la gestión del proyecto.

MODELO 3

- **Gráfica:** Dispersión

- **Hallazgos:**
  - Los puntos azules representan los valores reales de Project manager en función de Project organization.
  - Los puntos rojos corresponden a las predicciones del modelo.
  - Los datos reales se muestran dispersos, con agrupaciones y valores alejados de la tendencia central, lo que refleja que la relación entre variables no es completamente lineal.
  - Las predicciones forman una línea ascendente que sigue una tendencia positiva general, pero no captura la variabilidad y dispersión de los datos observados.
  - Esto confirma que la regresión lineal ofrece una aproximación simplificada, útil solo para identificar la tendencia global, pero con baja precisión predictiva frente a la complejidad real de los datos.
  -



MODELO 4

Variables independientes (X):

- BG: código de la unidad de negocio (Business Group).
- Project organization: organización del proyecto.
- Project manager: responsable de la gestión del proyecto.

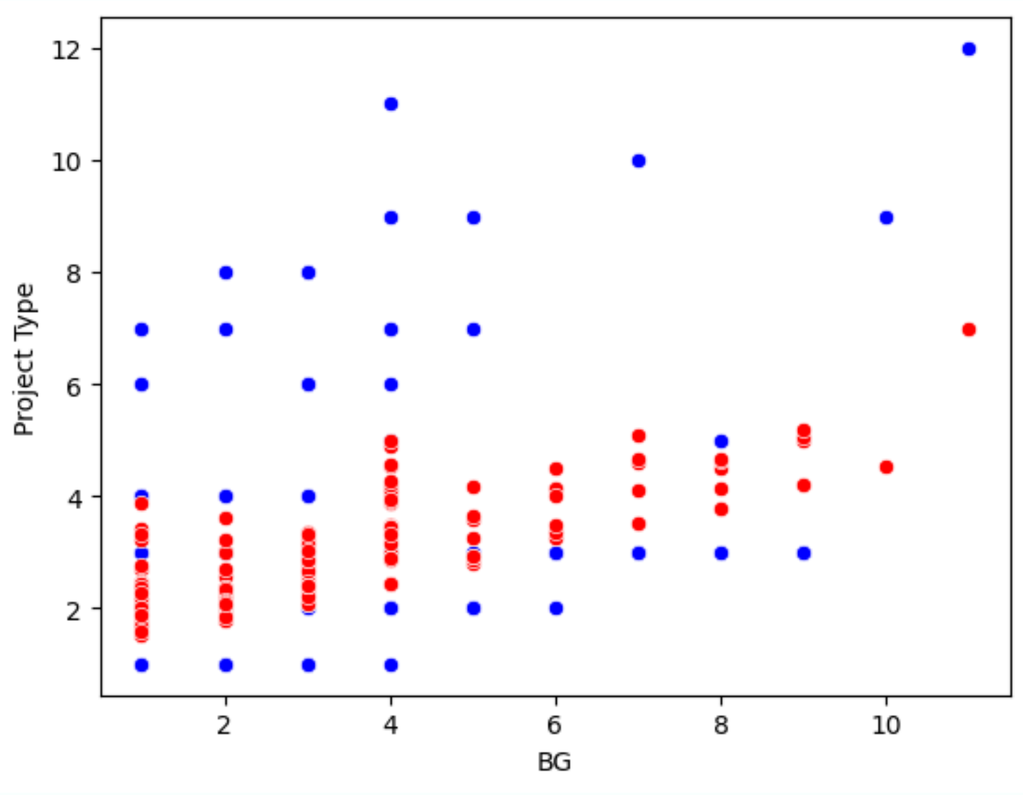
Variable dependiente (Y):

- Project Type: tipo de proyecto.

MODELO 4

- **Gráfica:** Dispersión

- **Hallazgos:**
  - Los puntos azules representan los valores reales de Project Type en función de BG.
  - Los puntos rojos corresponden a las predicciones generadas por el modelo.
  - Se observa que los valores reales se distribuyen en distintos niveles, con cierta dispersión vertical y sin un patrón lineal claro.
  - Las predicciones del modelo forman una franja ascendente relativamente uniforme, que simplifica la variabilidad de los datos reales.
  - Esto sugiere que la regresión lineal capta una tendencia positiva entre las variables, pero no refleja la complejidad ni la dispersión real de los valores, limitando su utilidad predictiva.



CONCLUSIONES

La regresión lineal ayudó a identificar patrones y correlaciones generales en ambos conjuntos de datos.

En conjunto, los resultados muestran que la regresión lineal es útil para detectar tendencias generales, pero no suficiente para predicciones precisas, lo que abre la necesidad de más variables y técnicas avanzadas para capturar la complejidad real.

Octubre 2030

MUCHAS
GRACIAS

Bruno Lago