



Tecnológico de Monterrey

Instituto Tecnológico y de Estudios Superiores de Monterrey Campus Puebla

Analítica de datos y herramientas de inteligencia artificial I (Gpo 101)

Actividad 2.1 (Regresión Lineal Simple y Múltiple)

Estudiantes: María Matanzo Hermoso | A01737554

Laisha Puentes Angulo | A012736397

Marco Cornejo Cornejo | A01276411

Jorge Alberto Cortes Sánchez | A01736236

Eduardo Torres Naredo | A01734935

Fecha: 3 de Octubre de 2025

Reporte de Hallazgos – Actividad 2.2 (FORVIA)

Introducción

El objetivo de esta actividad fue **analizar relaciones entre variables de proyectos de FORVIA** aplicando regresión lineal simple y múltiple. El análisis siguió estos pasos:

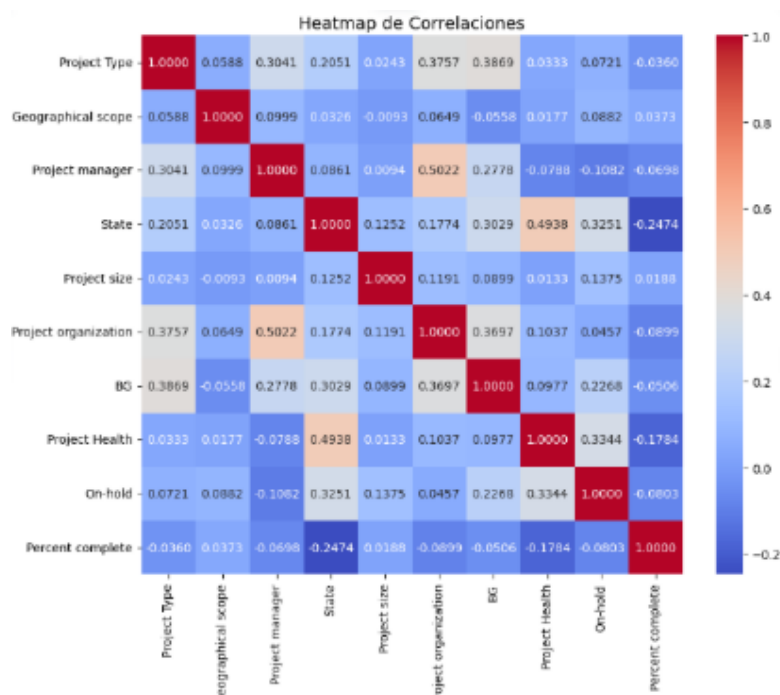
1. **Preparación de datos:** se transformaron variables categóricas en valores numéricos según su frecuencia de aparición (jerarquía en frecuencias).
2. **Correlaciones iniciales (Heatmap):** se identificaron los pares de variables más relacionados.
3. **Regresión lineal múltiple:** se construyeron modelos donde cada variable cuantitativa fue tratada como dependiente, explicada por varias independientes.
4. **Interpretación profesional:** se evaluó la capacidad de cada modelo para explicar o predecir comportamientos, usando R^2 y gráficas comparativas.

1. Preparación de datos

Variables convertidas a numéricas: **Project Type, Geographical scope, Project manager, State, Project size, Project organization, BG, Project Health, On-hold.**

Se obtuvo un DataFrame exclusivamente numérico, listo para análisis.

2. Correlaciones principales (Heatmap)



Los cinco pares de variables con mayor correlación fueron:

1. **Project manager – Project organization** → 0.502
2. **Project Health – State** → 0.494
3. **BG – Project Type** → 0.387
4. **Project Type – Project organization** → 0.376
5. **BG – Project organization** → 0.370

Esto muestra que **la forma de organizar y gestionar los proyectos está fuertemente asociada a quién los dirige**, y que **la salud del proyecto se relaciona con su estado**.

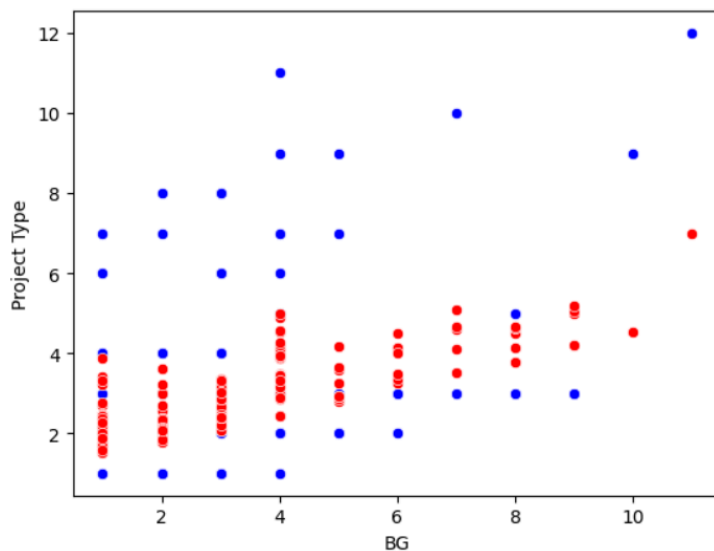
3. Modelos de Regresión Lineal Múltiple

Formato: Variables + Interpretación sencilla (puntos azules = reales, puntos rojos = predicciones) + R^2 .

Modelo 1 — Project Type

X: BG, Project organization, Project manager

Y: Project Type



Interpretación

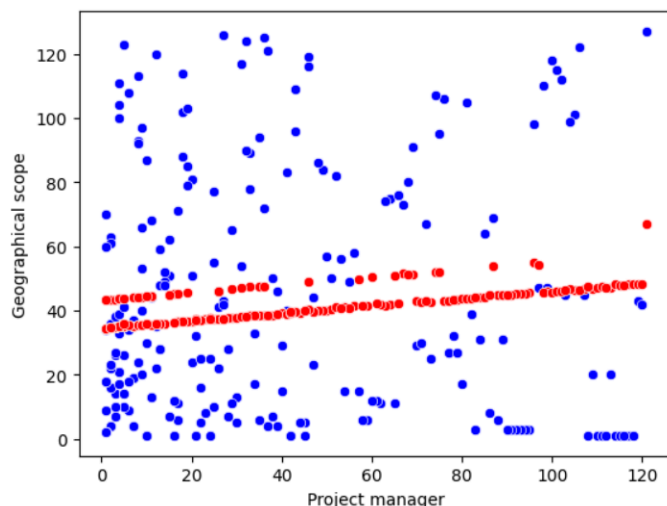
- **Puntos azules:** se observan **grupos** de tipos de proyecto asociados a ciertos **BG** y **organizaciones**; hay variabilidad dentro de cada grupo.
- **Puntos rojos:** siguen una **tendencia promedio** (línea/franja), pero **no alcanzan** a reproducir la **dispersión interna** de los grupos.
- El modelo **detecta** que “quién organiza” y el **BG** influyen en el tipo de proyecto, pero **no captura** diferencias finas entre prácticas locales.

Resultado: $R^2 \approx 22\%$ ($R \approx 0.47$) → **describe tendencia, predicción limitada.**

Modelo 2 — Geographical scope

X: Project manager, On-hold, Project organization

Y: Geographical scope



Interpretación

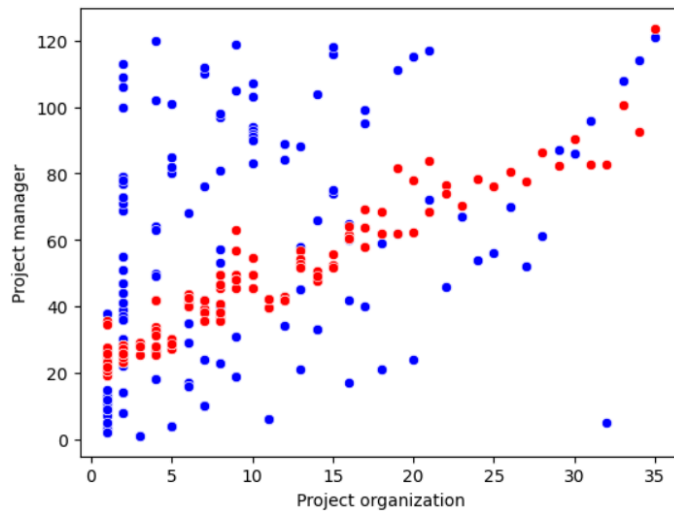
- **Puntos azules:** muy **dispersos**; el alcance geográfico aparece en **múltiples combinaciones** de manager/organización/pausa, **sin patrón claro**.
- **Puntos rojos:** casi una **línea plana**; el modelo **no encuentra estructura** utilizable.
- El alcance geográfico **no es linealmente explicable** con estas variables (posible influencia de políticas regionales, clientes, logística).

Resultado: $R^2 \approx 2\%$ ($R \approx 0.14$) → **no explicativo.**

Modelo 3 — Project manager

X: Project organization, Project Type, BG

Y: Project manager



Interpretación

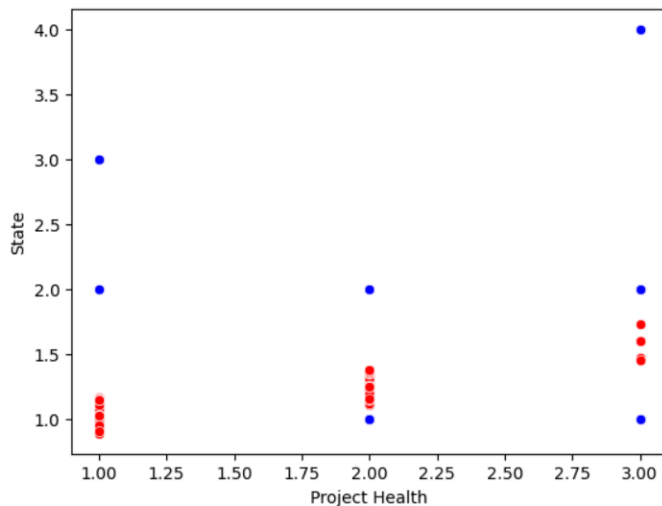
- **Puntos azules:** se aprecian **preferencias** (ciertos managers aparecen más en **algunas organizaciones y tipos**), pero hay **mucha dispersión**.
- **Puntos rojos:** siguen una **dirección general** (captan parte de esa preferencia), **sin precisión individual**.
- El modelo revela **alineaciones** entre manager–organización–BG, pero la **alta cardinalidad** de “manager” y la **dispersión** reducen la capacidad predictiva.

Resultado: $R^2 \approx 27\%$ ($R \approx 0.52$) → **útil para tendencias, no para asignación fina.**

Modelo 4 — State

X: Project Health, On-hold, Percent complete

Y: State



Interpretación

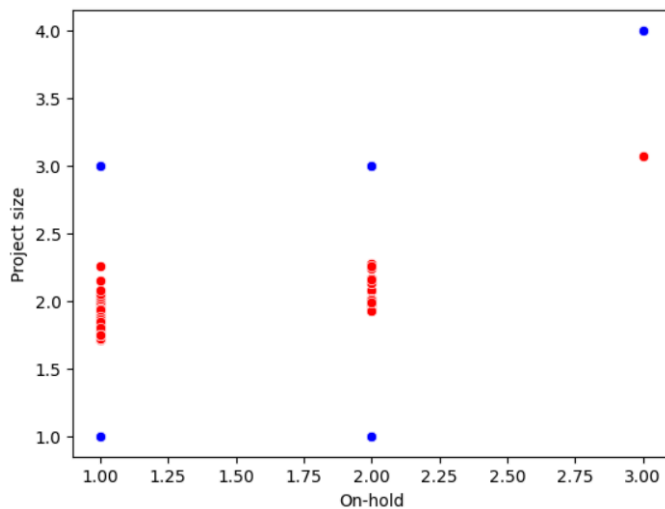
- **Puntos azules:** estados **mejores** suelen aparecer con **mayor salud y más avance**; se ven **estratos** (grados de estado) con variación interna.
- **Puntos rojos:** siguen esa **gradiente** (más salud/avance → mejor estado), pero **suavizan** las diferencias entre estratos.
- El estado se **relaciona** con salud/avance, pero hay **detalles operativos** (hitos, riesgos) que el modelo lineal **no contempla**.

Resultado: $R^2 \approx 30\%$ ($R \approx 0.55$) → **moderado, orienta pero no clasifica** con precisión.

Modelo 5 — Project size

X: On-hold, State, Project organization

Y: Project size



Interpretación

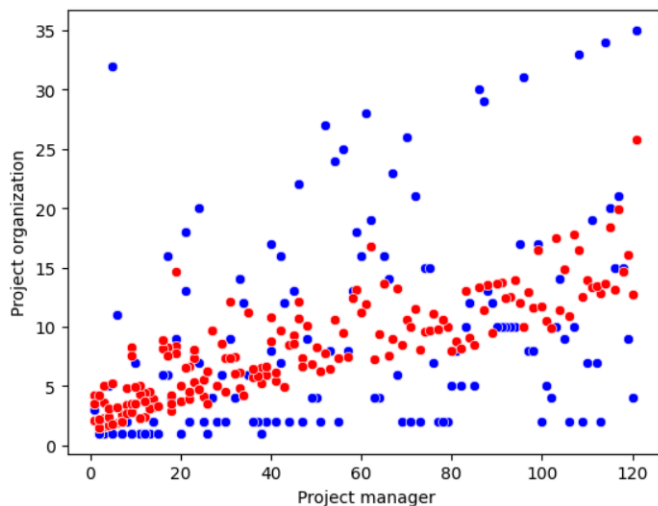
- **Puntos azules:** tamaños **pequeños/medianos/grandes** aparecen en **muchas combinaciones** de estado/organización/pausa; no hay **gradiente lineal**.
- **Puntos rojos:** se concentran en una **recta** que **no refleja** la **variabilidad real** de tamaños.
- El tamaño parece depender de **factores omitidos** (presupuesto, complejidad técnica, duración, alcance contractual).

Resultado: $R^2 \approx 4\%$ ($R \approx 0.19$) → **muy débil, no usable** para explicar tamaño.

Modelo 6 — Project organization

X: Project manager, Project Type, BG

Y: Project organization



Interpretación

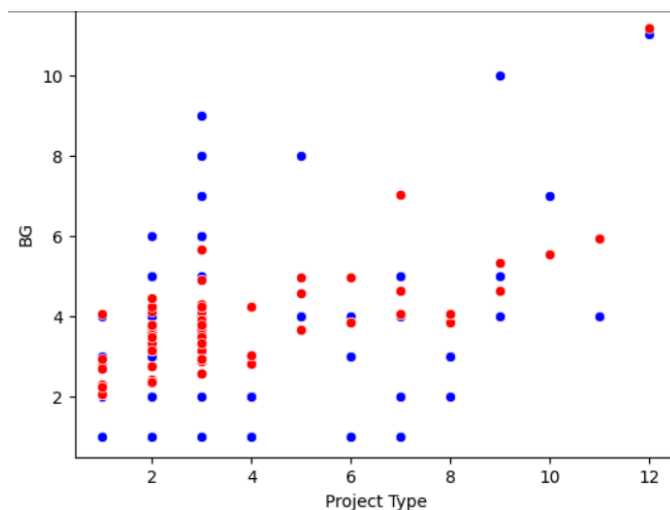
- **Puntos azules:** la organización se **asocia** con **BG** y **tipo**; hay **agrupamientos** visibles por familia de proyectos.
- **Puntos rojos:** capturan bien la **dirección** (la franja roja sigue los grupos azules), aunque con **simplificación**.
- El modelo es **coherente** para entender cómo se **distribuyen** los proyectos entre organizaciones, útil para **planificación**.

Resultado: $R^2 \approx 34\%$ ($R \approx 0.58$) → **de lo más sólido** del conjunto.

Modelo 7 — BG

X: Project Type, Project organization, State

Y: BG



Interpretación

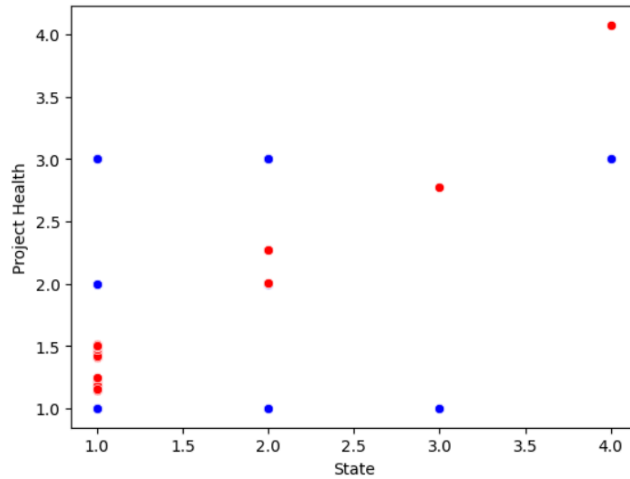
- **Puntos azules:** algunos **BG** aparecen preferentemente en **ciertos tipos y organizaciones**; aún así, hay **solapamientos**.
- **Puntos rojos:** reproducen la **tendencia** (BG alineado con tipo/organización), pero **no** toda la **variación**.
- Útil para ver **alineación estratégica** BG–tipo–organización; **limitado** para asignación determinística.

Resultado: $R^2 \approx 25\%$ ($R \approx 0.50$) → **moderado**.

Modelo 8 — Project Health

X: State, On-hold, Percent complete

Y: Project Health



Interpretación

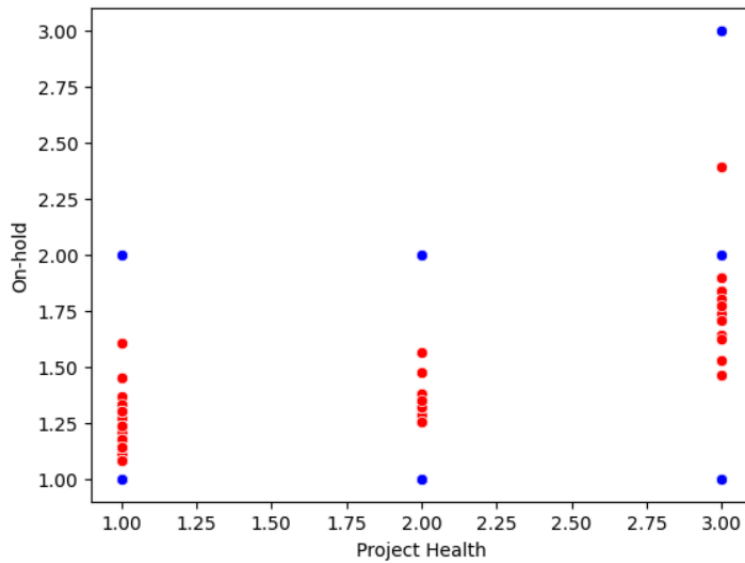
- **Puntos azules:** más avance y mejor estado → mejor salud; pausa (On-hold) → peor salud; el patrón es claro, con variación.
- **Puntos rojos:** siguen bien la **dirección esperada** (sube salud con avance/estado; baja con pausa).
- El modelo **explica** la **salud** a nivel global; para decisiones finas conviene agregar variables de **riesgo** y **recursos**.

Resultado: $R^2 \approx 28\%$ ($R \approx 0.53$) → **coherente** y **útil** para **monitoreo**.

Modelo 9 — On-hold

X: Project Health, State, BG

Y: On-hold



Interpretación

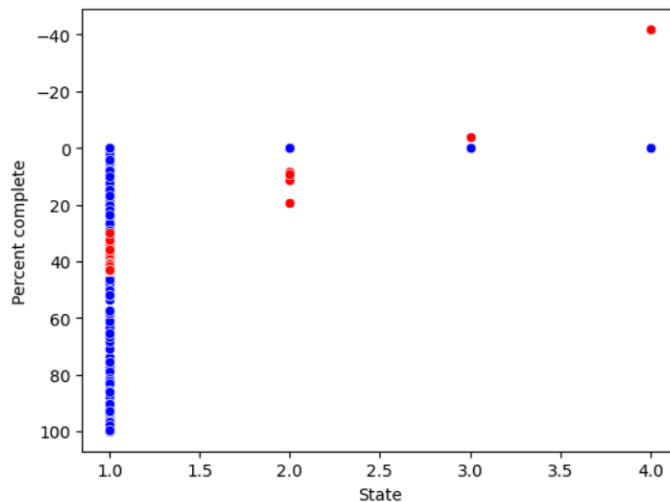
- **Puntos azules:** proyectos en pausa aparecen a lo largo de **muchos valores** de salud/estado; **no** hay **frontera clara**.
- **Puntos rojos:** muestran solo una **ligera** relación (peor salud → más probabilidad de pausa).
- “Pausa” parece depender de **choques externos** (bloqueos, dependencia de terceros, aprobaciones) **no** reflejados en el dataset.

Resultado: $R^2 \approx 17\%$ ($R \approx 0.41$) → **débil**, se requieren **otros factores**.

Modelo 10 — Percent complete

X: State, Project Health, Project organization

Y: Percent complete



Interpretación

- **Puntos azules:** mayor salud/estado se asocian con **más avance**, pero con **mucha dispersión** (proyectos que avanzan más/menos de lo esperado).
- **Puntos rojos:** siguen la **tendencia** (más salud/estado → más % completado), pero **no** capturan los **desvíos**.
- El avance está **parcialmente** explicado; faltan variables de **cronograma**, **recursos**, **alcance** y **bloqueadores**.

Resultado: $R^2 \approx 7\%$ ($R \approx 0.26$) → **bajo**, no confiable para pronóstico.

4. Conclusiones Generales de la Actividad 2.2

- Los modelos con mejor ajuste fueron **Project organization (34%)**, **State (30%)**, **Project manager (27%)** y **Project Health (28%)**, que muestran **tendencias moderadas**.
- Los más débiles: **Geographical scope (2%)**, **Project size (4%)**, y **Percent complete (7%)**, donde no hay relación lineal clara.
- El **Top-5 de correlaciones** muestra la importancia de la **gestión** (manager/organización) y la **salud del proyecto** como factores clave.

Conclusión

Aunque la regresión lineal múltiple permitió identificar algunas relaciones relevantes, los modelos no alcanzan valores altos de R^2 . Esto sugiere que para FORVIA sería más adecuado complementar el análisis con métodos no lineales (árboles, random forest) o incorporar más variables (tiempo, recursos asignados, complejidad del proyecto).