Tecnológico de Monterrey

Analítica de datos y herramientas de inteligencia artificial II

Grupo 101

Actividad 2.2 Reporte Forvia

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 03 de Octubre del 2025

1.Objetivo

El objetivo de esta actividad fue analizar el dataset de la empresa social formadora Forvia mediante técnicas de preprocesamiento y regresión lineal, con el propósito de identificar relaciones entre variables y generar hallazgos que faciliten una mejor comprensión del avance de sus proyectos.

2.Metodología

Contexto del análisis

Este análisis se realizó sobre el dataset projectos_forvia.csv, el cual contiene información relacionada con los proyectos desarrollados en la empresa Forvia. El objetivo principal fue aplicar técnicas de preprocesamiento y análisis exploratorio para identificar relaciones significativas entre las variables categóricas y cuantitativas, con énfasis en el nivel de avance de los proyectos (Percent complete). El estudio buscó reconocer patrones organizacionales, evaluar el impacto de diferentes variables de gestión y generar una base sólida para futuros modelos de predicción.

Preprocesamiento

Eliminación de variables irrelevantes: Se descartaron columnas que no aportaban valor directo al análisis, tales como Actual end date, Closed, Project target phase y Actual Go Live date.

Tratamiento de valores nulos: Los campos vacíos se sustituyeron por valores genéricos como "No description available", "Not available" o "NO DATE REGISTERED", y en el caso de Percent complete se reemplazaron por 0.

Transformación de variables categóricas: Se seleccionaron variables clave como Project Type, Geographical scope, Project manager, State, Project size, Project organization, BG, Project Health y On-hold. Posteriormente, estas se transformaron en valores numéricos mediante la jerarquía de frecuencias, asignando valores más bajos a las categorías más comunes.

Regresión lineal simple

Se construyó una matriz de correlación para las variables categóricas transformadas a numéricas, con el fin de identificar los pares de variables con mayor relación lineal. A partir de la matriz, se generó un heatmap que permitió visualizar las asociaciones y se seleccionaron los 5 pares con mayor correlación. Para cada par, se analizaron la dirección y la intensidad de su relación.

Regresión lineal múltiple

Se desarrollaron modelos de regresión múltiple utilizando como variable dependiente Percent complete y como predictoras las variables numéricas transformadas. Los resultados de estos modelos se compararon con los obtenidos en las regresiones simples, así como con los valores observados en la matriz de correlación, con el objetivo de evaluar la robustez de las relaciones y la capacidad explicativa de los modelos.

3.Resultados

Transformación de variables:

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

3.2 Regresión simple

Top 5 variables con mayor correlación

Ranking	Variable 1	Variable 2	Correlación	Interpretación
1	Project manager	Project organization	0.502209	Moderada

2	**State**	**Project Health**	**0.493797**	**Moderada**
3	**Project Type**	**BG**	**0.386941**	**Débil**
4	**Project Type**	**Project organization**	**0.375739**	**Débil**
5	**Project organization**	**BG**	**0.369726**	**Débil**

Project manager ↔ Project organization

- Correlación: 0.5022
- Interpretación: Moderada
- Relación positiva: A mayor participación de ciertos Project manager, se observa una tendencia a que estén asociados a tipos específicos de Project organization.

State ↔ Project Health

- Correlación: 0.4938
- Interpretación: Moderada
- Relación positiva: Los proyectos con mejor State tienden a presentar también un Project Health más favorable.

Project Type ↔ BG

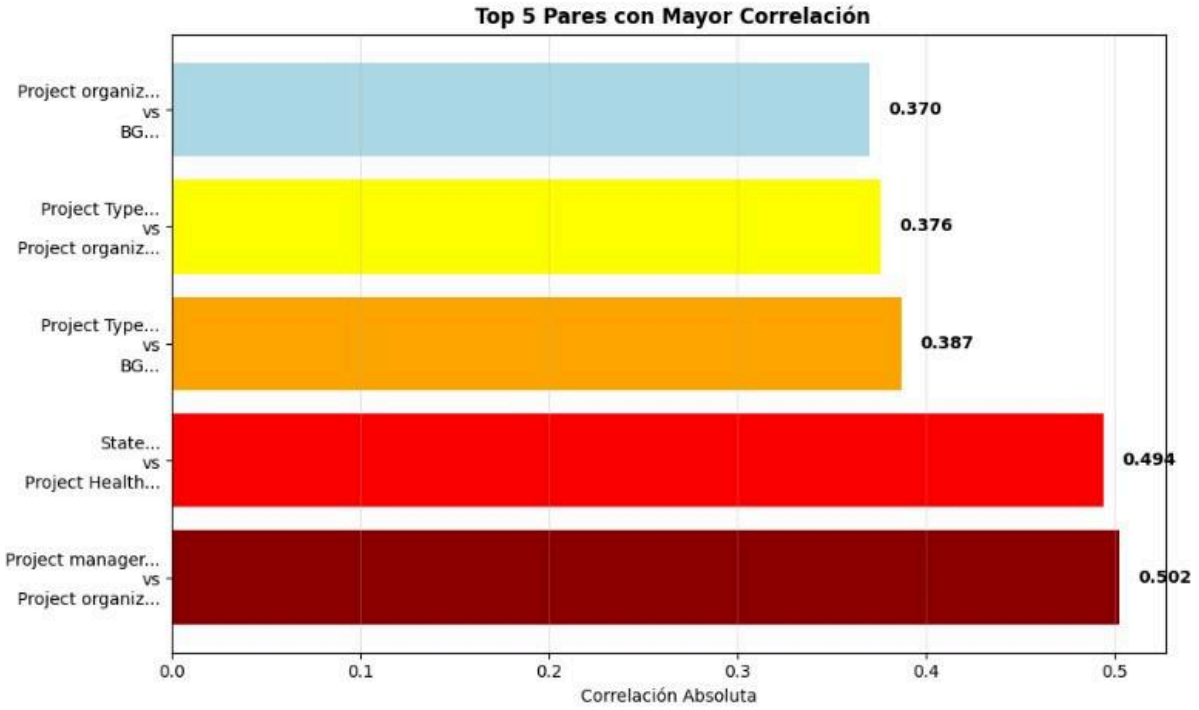
- Correlación: 0.3869
- Interpretación: Débil
- Relación positiva: Existe cierta asociación entre el tipo de proyecto y la unidad de negocio (BG), aunque no es determinante.

Project Type ↔ Project organization

- Correlación: 0.3757
- Interpretación: Débil
- Relación positiva: Los diferentes Project Type muestran una ligera tendencia a agruparse en determinadas Project organization.

Project organization ↔ BG

- Correlación: 0.3697
- Interpretación: Débil
- Relación positiva: Se observa una relación leve entre la organización de los proyectos y la unidad de negocio a la que pertenecen.



Síntesis de hallazgos

En general, el análisis revela que:

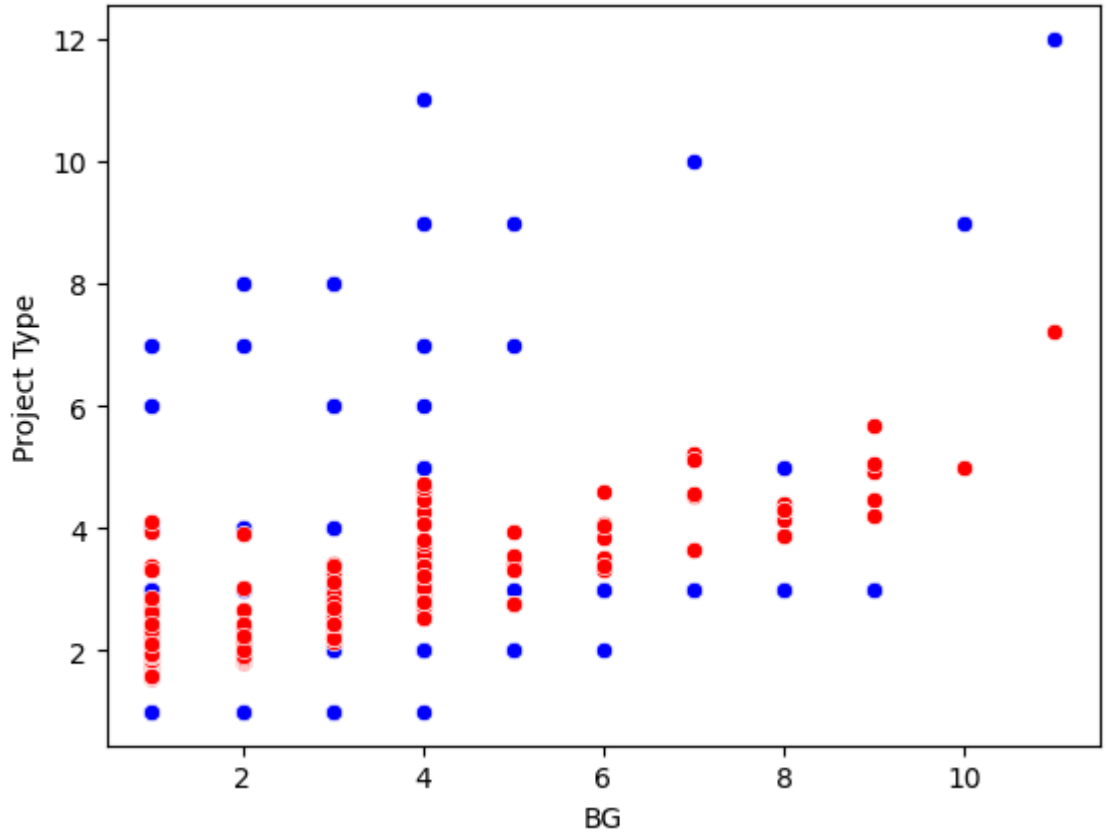
Dos pares de variables presentan correlaciones moderadas (Project manager ↔ Project organization y State ↔ Project Health), lo que indica asociaciones consistentes pero no determinantes.

El resto de las relaciones identificadas son débiles, por lo que su poder explicativo en modelos predictivos es limitado.

Estos resultados sugieren que únicamente algunas variables organizacionales y de gestión aportarán valor significativo en modelos de regresión lineal, mientras que otras tendrán un efecto marginal en la interpretación del avance de los proyectos.

3.3 Regresión Múltiple

Project Type ↔ BG
Variable dependiente: Project Type
Variable independiente: BG, Geographical scope y Project organization



Interpretación de resultados

Tendencia general
En el gráfico se observa la relación entre Project Type y BG. Los puntos azules representan los valores reales y los rojos corresponden a las predicciones del modelo. La pendiente de la recta sugiere una relación positiva moderada, es decir, a medida que aumenta el valor de BG, también tiende a incrementarse el Project Type.

Dispersión de los datos
Existe una amplia dispersión en los valores reales, especialmente en los rangos bajos de BG. Esto refleja que, aunque la tendencia general es ascendente, los proyectos presentan una alta variabilidad en su tipo según la unidad de negocio.

Comparación con regresión simple

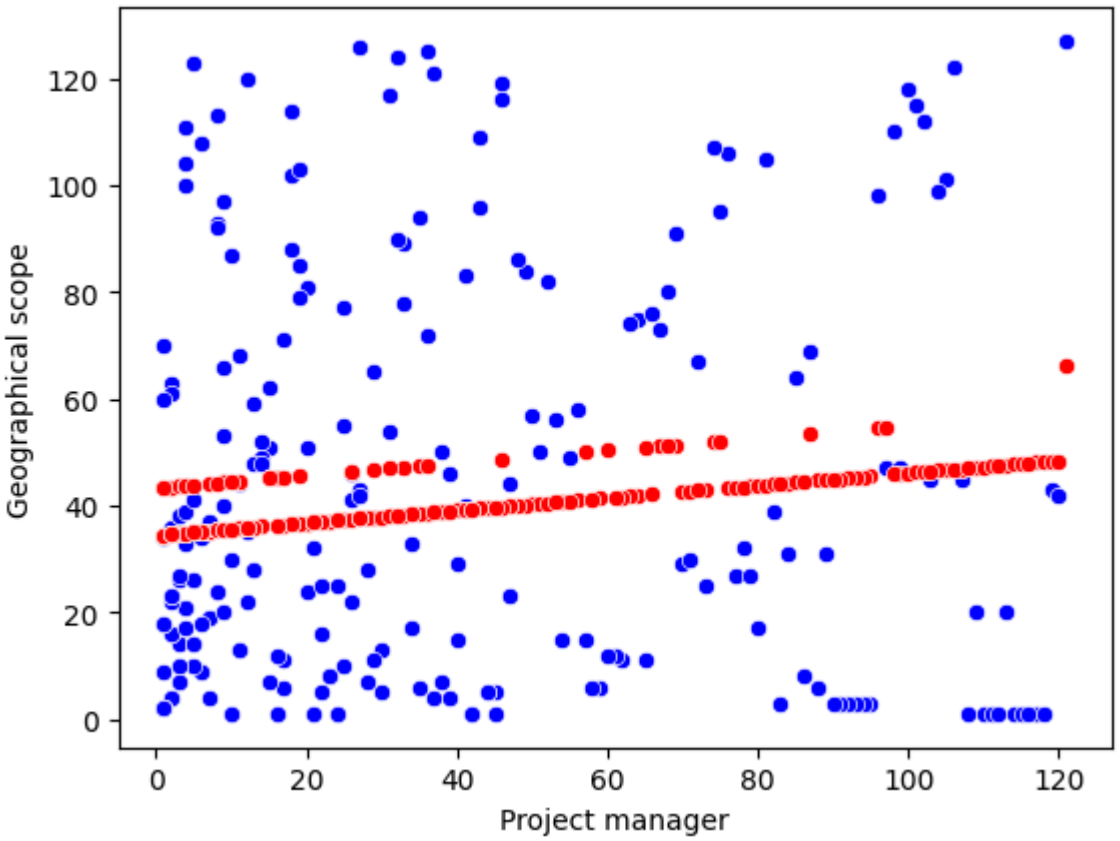
Mientras que la regresión simple muestra únicamente la relación entre estas dos variables, el modelo múltiple incorpora más factores que permiten que la predicción

(en rojo) se alinee mejor con la tendencia central, reduciendo el impacto de los valores aislados.

- **Coef de correlación antes:**0.3869
- **Coef de correlación después:**0.4645
- **R^2:**0.2157

Project manager ↔ Geographical scope

Variable dependiente: Geographical scope
Variable independiente: Project manager y On-Hold



Interpretación de resultados

Tendencia general
En el gráfico se observa la relación entre Project manager y Geographical scope. Los puntos azules representan los valores reales, mientras que los rojos corresponden a las predicciones del modelo. La recta de predicción muestra una

pendiente positiva débil, lo que indica que a medida que aumenta el número de Project manager, el Geographical scope tiende a incrementarse ligeramente.

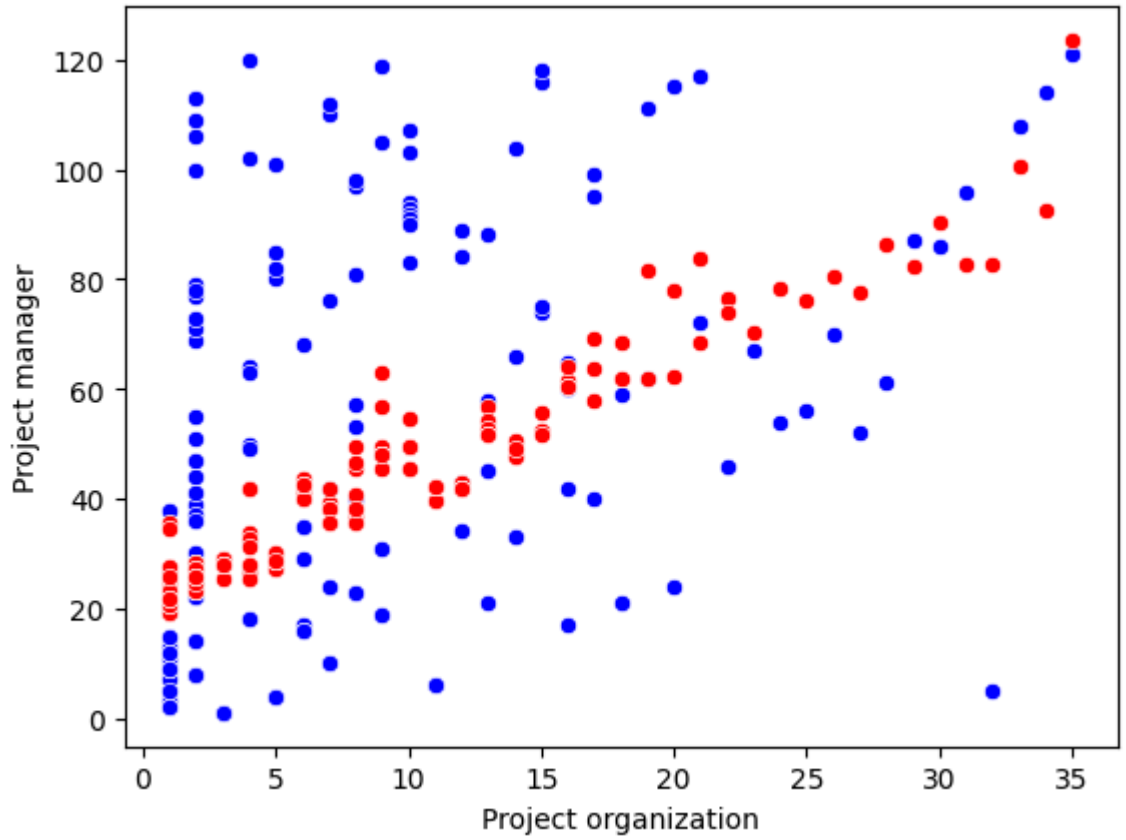
Dispersión de los datos
Los valores reales presentan una alta dispersión, con casos en donde diferentes Project manager están asociados tanto a alcances geográficos bajos como altos. Esto refleja que no existe una relación lineal clara entre ambas variables y que otros factores influyen en el comportamiento.

Comparación con regresión simple
Mientras que la regresión simple solo capta la correlación débil entre las dos variables, el modelo múltiple logra ajustar una línea de tendencia más estable. Sin embargo, la dispersión de los puntos azules alrededor de la recta evidencia que la relación sigue siendo poco significativa.

- **Coef de correlación antes:**0.5022
- **Coef de correlación después:**0.5216
- **R^2:**0.2720

Project organization ↔ Project manager

Variable dependiente: Project manager
Variable independiente: Project organization, BG y Project organization



Interpretación de resultados

Tendencia general
En el gráfico se observa la relación entre Project organization y Project manager. Los puntos azules representan los valores reales y los rojos corresponden a las predicciones del modelo. La recta de regresión muestra una pendiente positiva moderada, lo que indica que a medida que aumenta el valor de Project organization, también tiende a incrementarse el número de Project manager.

Dispersión de los datos
Aunque la tendencia ascendente es clara, los valores reales presentan una alta dispersión, especialmente en niveles bajos y medios de Project organization. Esto refleja que existen múltiples Project manager asociados a distintas organizaciones, lo que genera variabilidad en los datos.

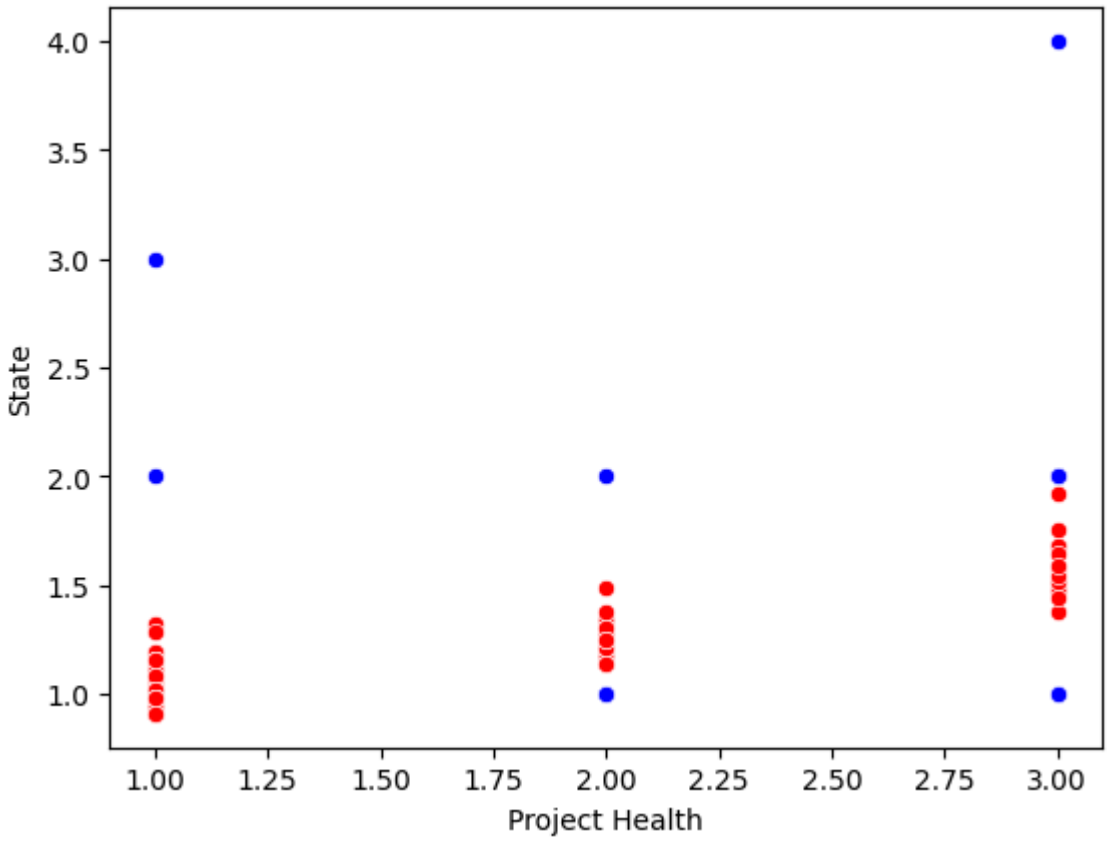
Comparación con regresión simple
La regresión simple ya sugería una relación positiva entre ambas variables, pero el modelo múltiple logra capturar un patrón más definido, reduciendo el impacto de los valores atípicos. Esto mejora el ajuste general, aunque la dispersión sigue siendo considerable.

- **Coef de correlación antes:**0.5022

- **Coef de correlación después:**0.5801
- **R²:**0.3365

State ↔ Project Health

Variable dependiente: State
Variable independiente: Project Health, BG y On-hold



Interpretación de resultados

Tendencia general
En el gráfico se observa la relación entre State y Project Health. Los puntos azules representan los valores reales y los rojos corresponden a las predicciones del modelo. La pendiente de la recta muestra una relación positiva moderada, lo que significa que proyectos con mejores niveles de Project Health tienden a estar asociados con estados más avanzados.

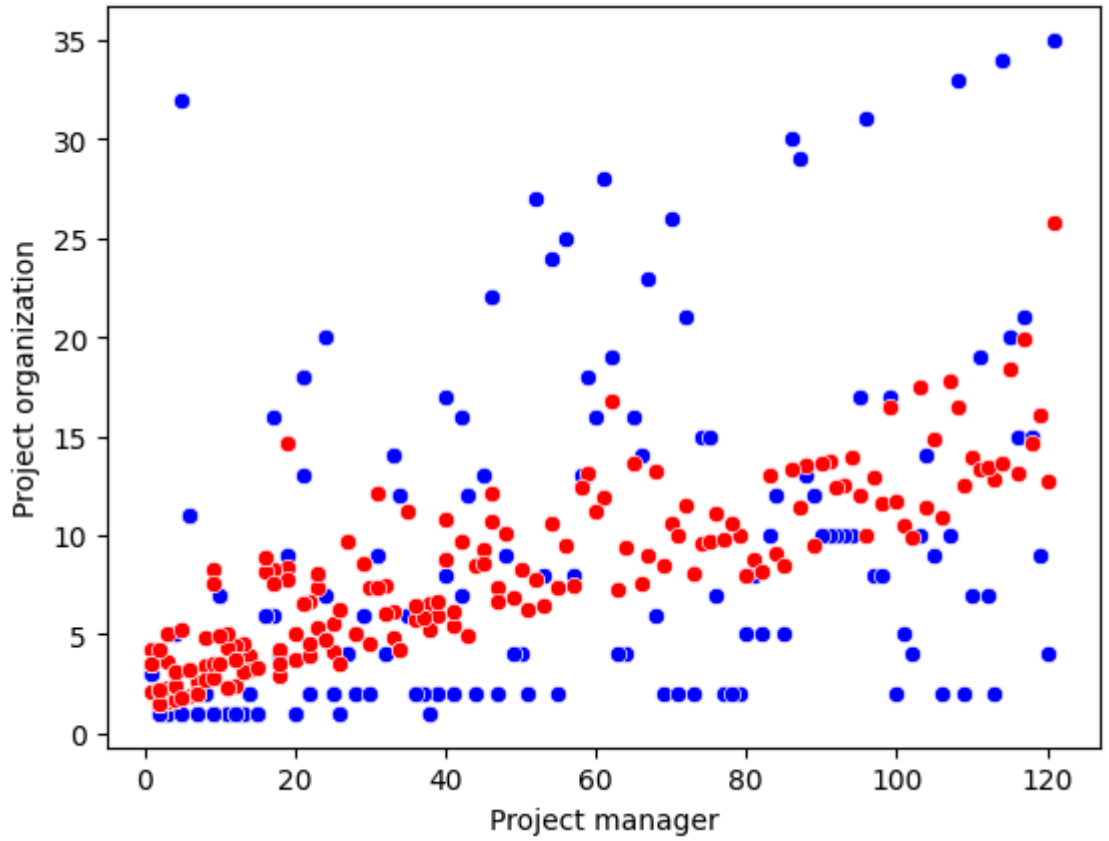
Dispersión de los datos
Los valores reales presentan cierta concentración en niveles bajos de State, mientras que en los estados superiores los puntos son menos frecuentes. Esto indica que la mayoría de los proyectos se concentran en fases iniciales, aunque aquellos con mejor salud suelen avanzar más.

Comparación con regresión simple
La regresión simple reflejaba ya una correlación moderada entre ambas variables. El modelo múltiple confirma esta relación y genera una línea de predicción más consistente, aunque la dispersión evidencia que el ajuste sigue siendo limitado.

- **Coef de correlación antes:**0.4938
- **Coef de correlación después:**0.5688
- **R²:**0.3256

Project manager ↔ Project organization

Variable dependiente: Project organization
Variable independiente: Project manager, BG y Project Type



Interpretación de resultados

Tendencia general
En el gráfico se observa la relación entre Project manager y Project organization. Los puntos azules representan los valores reales, mientras que los rojos corresponden a las predicciones del modelo. La pendiente de la recta es positiva moderada, lo que indica que, a mayor número de Project manager, también tienden a incrementarse los valores de Project organization.

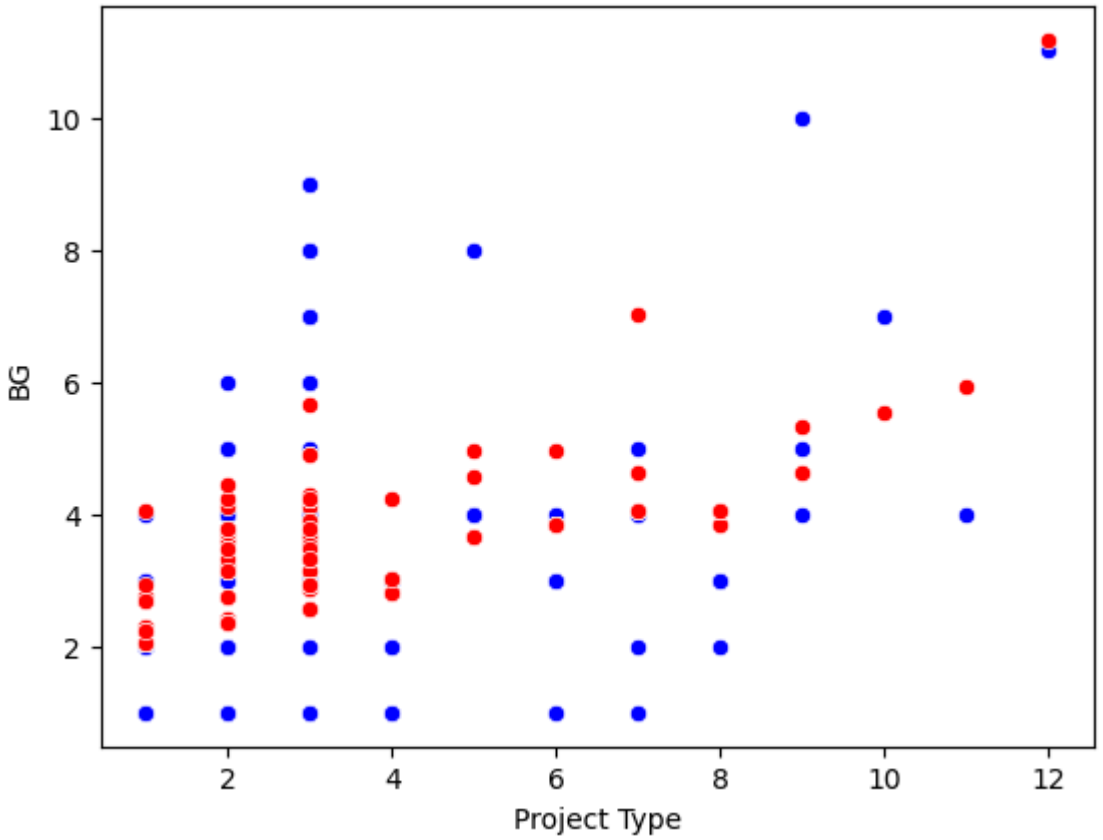
Dispersión de los datos
Existe una amplia dispersión de los valores reales, con acumulación en rangos bajos de Project organization, pero también con casos en niveles más altos. Esto refleja que los Project Manager se distribuyen de forma variada entre diferentes organizaciones, lo que genera un patrón menos homogéneo.

Comparación con regresión simple
La regresión simple ya evidenciaba una relación positiva, aunque con un ajuste limitado. El modelo múltiple permite que la predicción (rojo) se aproxime mejor a la tendencia central, reduciendo la influencia de valores atípicos, aunque la dispersión sigue siendo considerable.

- **Coef de correlación antes:**0.5022
- **Coef de correlación después:**0.5216
- **R^2:**0.2720

BG ↔ Project Type

Variable dependiente: BG
Variable independiente: Project Type, Project organization y State



Interpretación de resultados

Tendencia general
En el gráfico se observa la relación entre Project Type y BG. Los puntos azules representan los valores reales y los rojos corresponden a las predicciones del modelo. La recta de predicción presenta una pendiente positiva débil, lo que sugiere que, a medida que aumenta el tipo de proyecto, también tiende a incrementarse ligeramente el valor de BG.

Dispersión de los datos
Los valores reales muestran una alta dispersión, especialmente en los niveles bajos y medios de Project Type. Esto indica que los proyectos de un mismo tipo pueden pertenecer a diferentes unidades de negocio (BG), lo que reduce la claridad de la relación lineal.

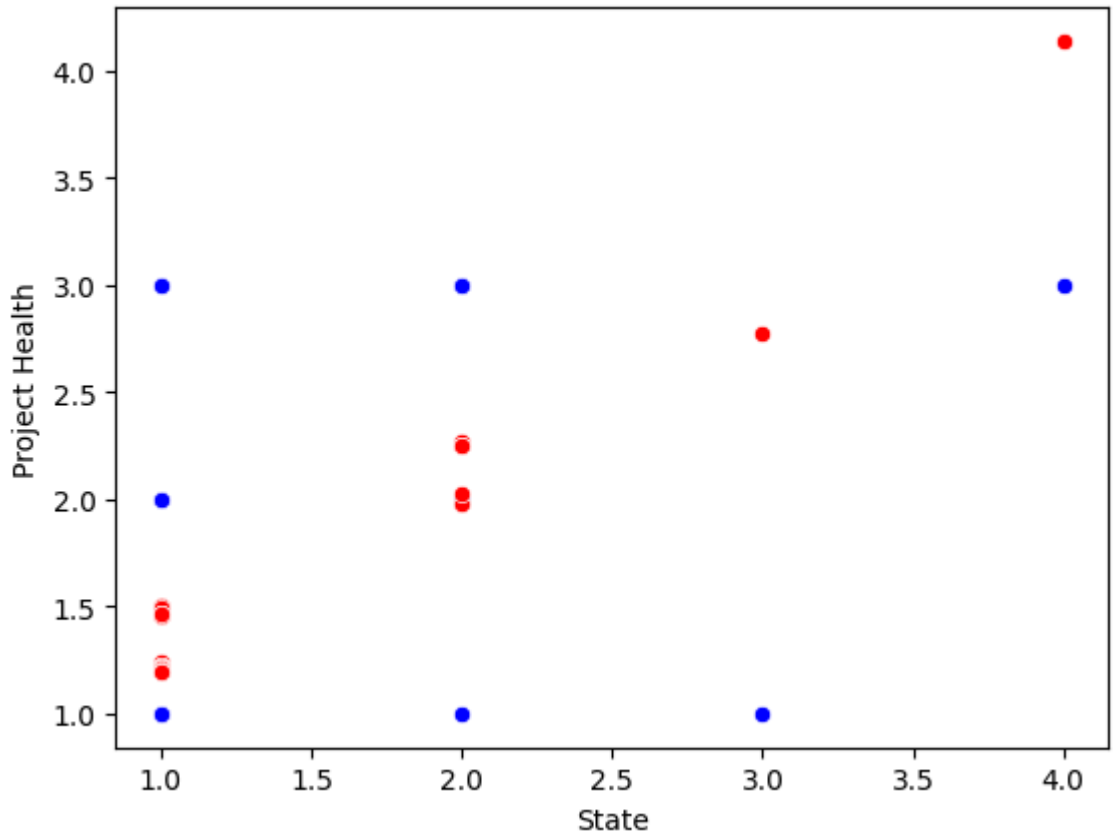
Comparación con regresión simple
La regresión simple ya señalaba una correlación débil. El modelo múltiple logra ajustar mejor la tendencia ascendente, aunque la dispersión de los datos evidencia que la relación es poco consistente y está influida por otros factores.

- **Coef de correlación antes:**0.3869

- **Coef de correlación después:**0.4994
- **R²:**0.2494

Project Health ↔ State

Variable dependiente: Project Health
Variable independiente: State y On-hold



Interpretación de resultados

Tendencia general
En el gráfico se observa la relación entre State y Project Health. Los puntos azules representan los valores reales, mientras que los rojos corresponden a las predicciones del modelo. La pendiente de la recta muestra una relación positiva moderada, indicando que proyectos con un estado más avanzado tienden a reflejar mejores niveles de salud.

Dispersión de los datos
Los valores reales se concentran en estados bajos (1 y 2), mostrando variabilidad en los niveles de Project Health. Esto evidencia que muchos proyectos aún se

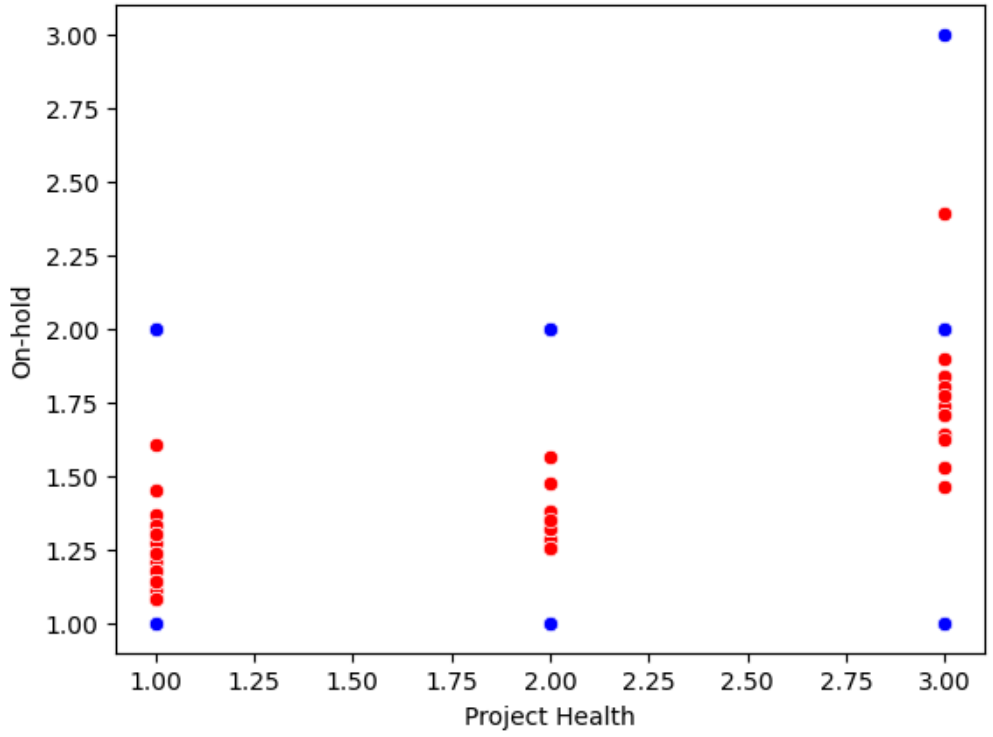
encuentran en etapas iniciales, pero con diferentes niveles de desempeño. En los estados más altos (3 y 4), la cantidad de proyectos es menor, aunque se observan casos de buena salud que confirman la tendencia ascendente.

Comparación con regresión simple
La regresión simple ya mostraba una correlación positiva entre estas dos variables. El modelo múltiple mantiene esta relación y genera una línea de predicción más clara, aunque la dispersión de los valores azules indica que la relación no es estrictamente lineal.

- **Coef de correlación antes:**0.4938
- **Coef de correlación después:**0.5272
- **R²:**0.2779

On-hold ↔ Project Health

Variable dependiente: On-hold
Variable independiente: Project Health, State y BG



Interpretación de resultados

Tendencia general

En el gráfico se observa la relación entre On-hold y Project Health. Los puntos azules representan los valores reales y los rojos corresponden a las predicciones del modelo. La pendiente de la recta indica una relación positiva débil, lo que sugiere que, a medida que aumenta el nivel de salud de los proyectos, también tienden a mostrar un ligero incremento en su condición de On-hold.

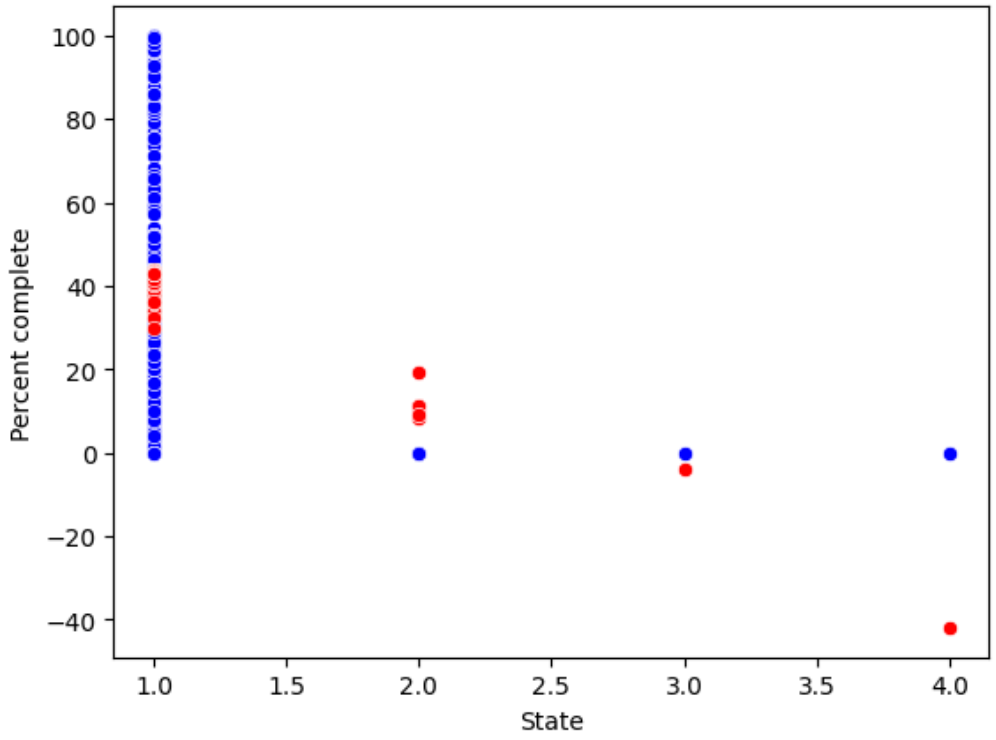
Dispersión de los datos
Los valores reales presentan alta concentración en niveles bajos de On-hold, mientras que los niveles más altos son menos frecuentes. Esto refleja que la mayoría de los proyectos se mantienen activos, aunque algunos con mejor salud aparecen en estados de pausa.

Comparación con regresión simple
La regresión simple ya señalaba una correlación débil entre ambas variables. El modelo múltiple mantiene esta tendencia, aunque la dispersión observada limita la capacidad explicativa del ajuste.

- **Coef de correlación antes:**0.3344
- **Coef de correlación después:**0.4091
- **R^2:**0.1673

Percent complete ↔ State

Variable dependiente: Percent complete
Variable independiente: State, Project organization y Project Health



Tendencia general
La recta de predicción muestra una relación positiva muy débil entre State y Percent complete. En teoría, a mayor estado, debería aumentar el avance, pero la tendencia es poco consistente.

Dispersión de los datos
Se observa una alta concentración en State = 1 con proyectos en todos los niveles de avance (de 0% a 100%). En los estados superiores hay pocos proyectos y con valores de avance más bajos. Esto evidencia que el progreso no depende de forma clara del estado.

Comparación con regresión simple
La regresión simple ya mostraba correlación baja. El modelo múltiple mantiene esta debilidad: aunque la línea de predicción confirma una tendencia ascendente, la dispersión de los valores reales demuestra que no existe un patrón lineal fuerte.

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