Especialización en Ciencia de Datos

Actividades

Estudiante: Yoseph Barrera
Profesor(a): Silvia Salinas Ayaviri
Fecha: 16 de febrero de 2026

Proyecto 4: Construcción de un panel público (WDI) y análisis exploratorio con regresión

Contexto y cambio de enfoque

El enunciado original del proyecto apuntaba a trabajar con una base asociada a *comercio*. Sin embargo, en el material de referencia no se disponía de un dataset público directo y listo para usar con ese foco específico. Por lo mismo, se ajustó el enfoque hacia una alternativa equivalente en espíritu y plenamente reproducible: construir un panel país–año con datos públicos del *World Development Indicators* (WDI) del Banco Mundial, y aplicar un flujo completo de ciencia de datos (descarga → limpieza → exploración → modelamiento).

Objetivo

Construir un dataset panel país–año (2010–2023) a partir de indicadores WDI y:

- describir estructura, tipos de variables, faltantes y outliers;
- explorar relaciones bivariadas y correlaciones;
- estimar un modelo de regresión lineal (OLS) para explicar el logaritmo del PIB per cápita.

Datos y fuente

Fuente de datos

Se utilizó la API oficial del Banco Mundial (*World Bank API*) para descargar indicadores WDI para "todos los países" (`country/all`) y el rango 2010–2023. El resultado es un panel con observaciones por país y año, identificado por el código ISO3 (`iso3c`).

Variables

Se definió como variable objetivo el PIB per cápita en US$ corrientes y como explicativas un conjunto de indicadores macro y estructurales:

Nombre corto	Indicador WDI
gdp_pc	NY.GDP.PCAP.CD (PIB per cápita, US$ corrientes)
life_exp	SP.DYN.LE00.IN (Esperanza de vida al nacer, años)
urban_pct	SP.URB.TOTL.IN.ZS (Población urbana, % del total)
co2_pc	EN.GHG.CO2.PC.CE.AR5 (Emisiones de CO_2 per cápita, AR5)
unemp	SL.UEM.TOTL.ZS (Desempleo, % fuerza laboral)
infl	FP.CPI.TOTL.ZG (Inflación, IPC % anual)
trade_gdp	NE.TRD.GNFS.ZS (Comercio, % del PIB)
internet	IT.NET.USER.ZS (Usuarios de internet, % población)

Preparación y limpieza

Construcción del panel

Para cada indicador se descargaron las observaciones país–año y luego se unieron mediante `merge` por `(iso3c, year)`. Se mantuvieron únicamente códigos ISO3 válidos (largo 3) y se eliminó cualquier registro sin `iso3c`.

Manejo de faltantes

El panel final contiene 3568 filas y 11 columnas (incluyendo `country` y `year`). La variable objetivo `gdp_pc` no presenta faltantes (se eliminan filas sin objetivo). Las explicativas sí tienen faltantes, en especial:

- `trade_gdp`: 513 nulos (≈14.38 %)
- `internet`: 487 nulos (≈13.65 %)
- `unemp`: 383 nulos (≈10.73 %)
- `infl`: 371 nulos (≈10.40 %)
- `co2_pc`: 177 nulos (≈4.96 %)

Transformación

Dado el fuerte sesgo a la derecha de `gdp_pc`, se creó `log_gdp_pc` para modelamiento:

$$\texttt{log\_gdp\_pc} = \log(\texttt{gdp\_pc})$$

Esta transformación mejora la interpretabilidad (cambios porcentuales aproximados) y reduce el impacto de valores extremos.

Productos guardados

- `outputs/data/wdi_dataset.csv`: panel base.
- `outputs/data/wdi_dataset_con_log.csv`: incluye `log_gdp_pc`.
- `outputs/figures/`: histogramas, boxplots, heatmaps y diagnósticos del modelo.

Análisis exploratorio (EDA)

Distribuciones y outliers

Las distribuciones muestran asimetrías relevantes (en particular `gdp_pc`, `infl` y `trade_gdp`). La regla IQR identificó outliers en varias variables, lo que refuerza el uso de transformaciones (como el log) y la necesidad de diagnósticos del modelo.

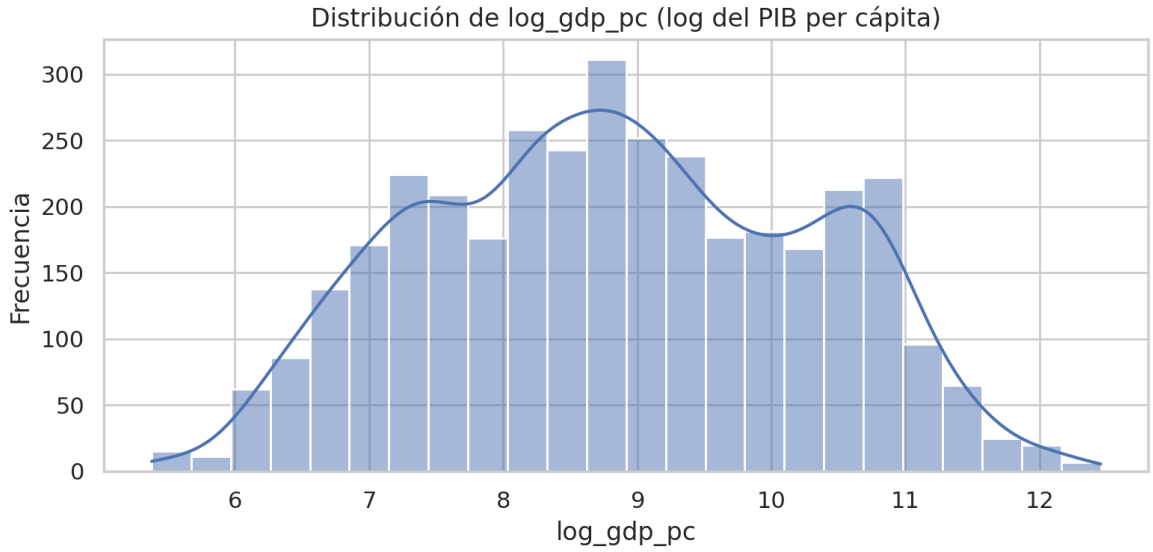


Figura 1: Distribución de `log_gdp_pc`.

Correlaciones

La matriz de correlación (Pearson) muestra asociaciones fuertes entre `log_gdp_pc` y: `life_exp` ($\approx$ 0,84), `internet` ($\approx$ 0,84), `urban_pct` ($\approx$ 0,69), `co2_pc` ($\approx$ 0,43). Por legibilidad, se reporta la matriz sin `year`, ya que el tiempo puede inducir correlaciones espurias.

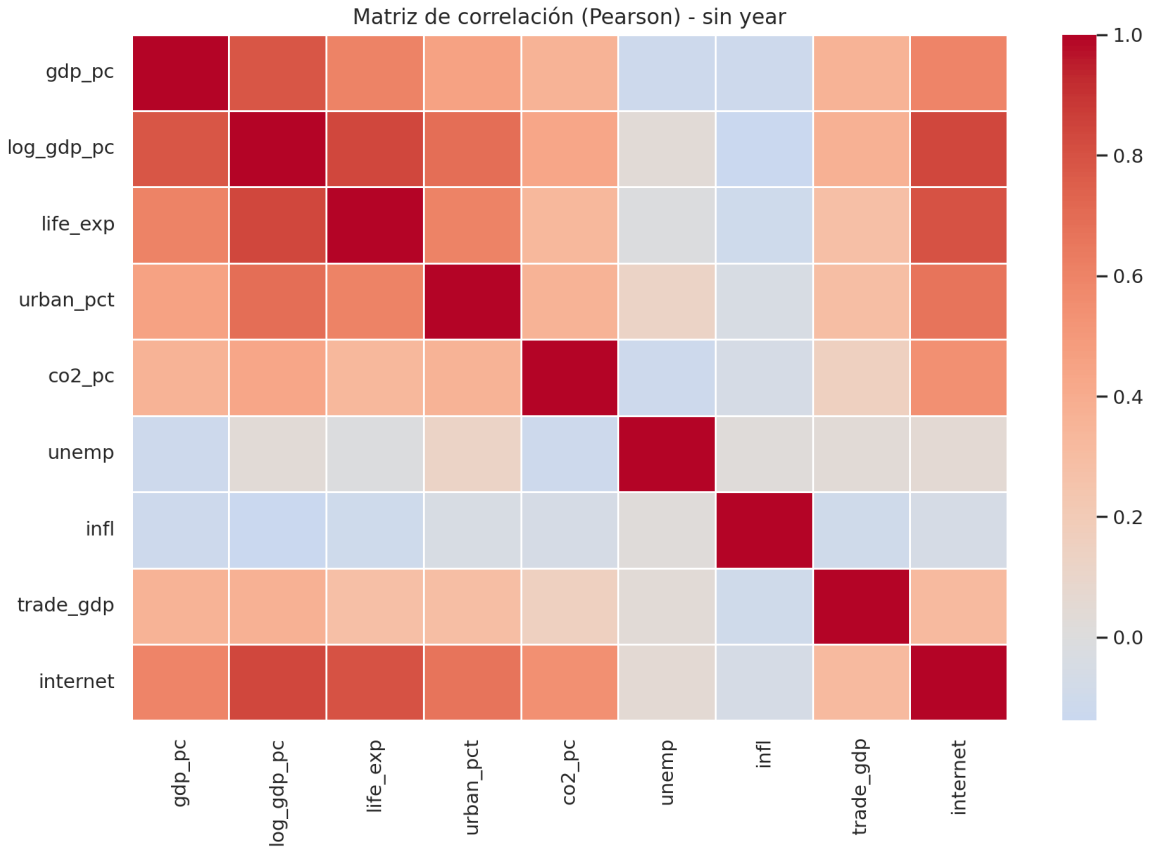
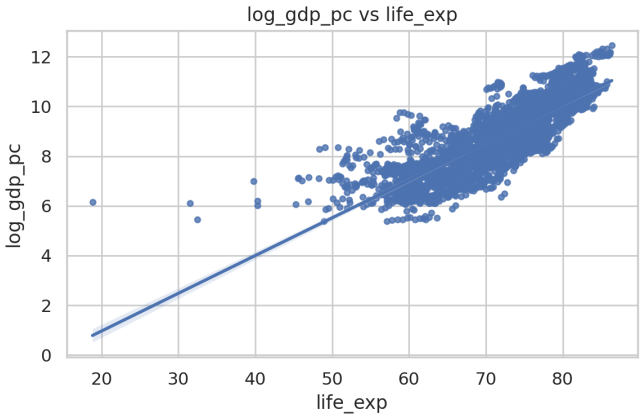


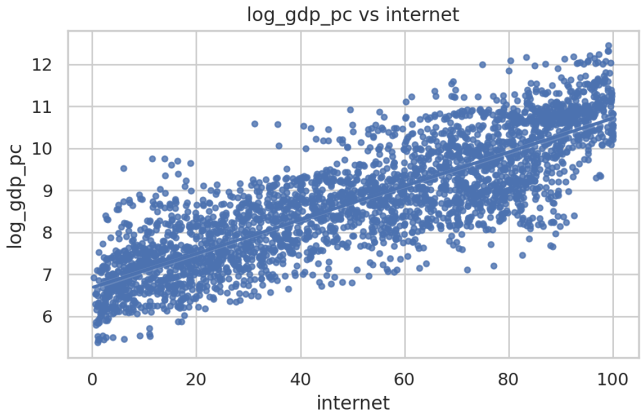
Figura 2: Matriz de correlación (Pearson) sin `year`.

Relaciones bivariadas con la variable objetivo

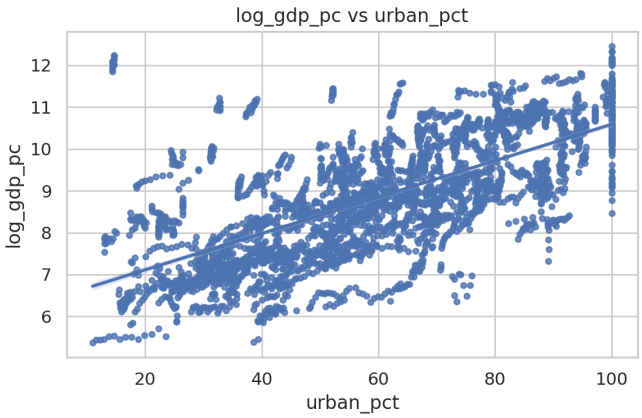
Se graficaron `log_gdp_pc` versus `life_exp`, `internet`, `urban_pct` y `co2_pc`. En general se observa una relación positiva clara para `life_exp` e `internet`; en `co2_pc` aparecen valores extremos que generan una nube más heterogénea.



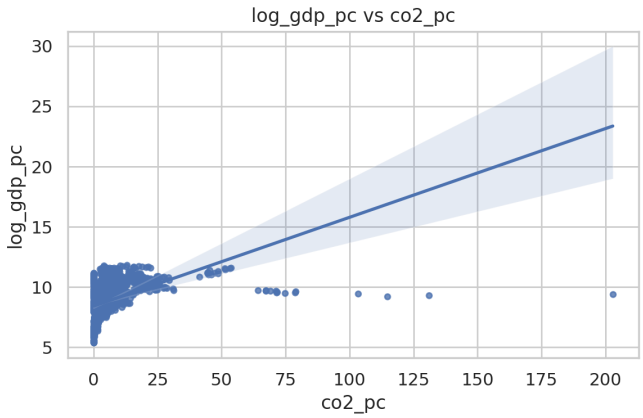
(a) `life_exp`



(b) `internet`



(c) `urban_pct`



(d) `co2_pc`

Figura 3: Relación bivariada entre `log_gdp_pc` y covariables seleccionadas.

Modelo: Regresión lineal (OLS)

Especificación

Se estimó el siguiente modelo:

$$\texttt{log\_gdp\_pc}_{it} = \beta_0 + \beta_1 \texttt{life\_exp}_{it} + \beta_2 \texttt{internet}_{it} + \beta_3 \texttt{urban\_pct}_{it} + \beta_4 \texttt{co2\_pc}_{it} + \beta_5 \texttt{trade\_gdp}_{it} + \beta_6 \texttt{infl}_{it} + \beta_7 \texttt{unem}$$

donde i es país y t es año.

Muestra usada

- Observaciones totales del panel: 3568.
- Observaciones con todas las variables del modelo no nulas: 2441.

- Se aplicó partición entrenamiento/prueba, quedando 1830 observaciones para estimación (train) y el resto para evaluación (test).

Resultados principales

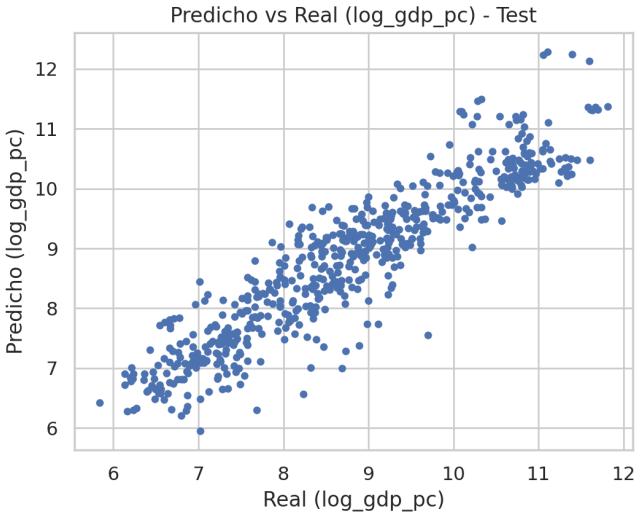
El ajuste del modelo es alto: $R^2 \approx 0{,}852$ (train) y $R^2 \approx 0{,}849$ (test). En la estimación, `life_exp`, `internet`, `urban_pct`, `co2_pc`, `trade_gdp` resultan positivas y estadísticamente significativas; `infl` es negativa y significativa. `unemp` no es estadísticamente significativa al 5 % (p $\approx$ 0.36).

Métricas en test:

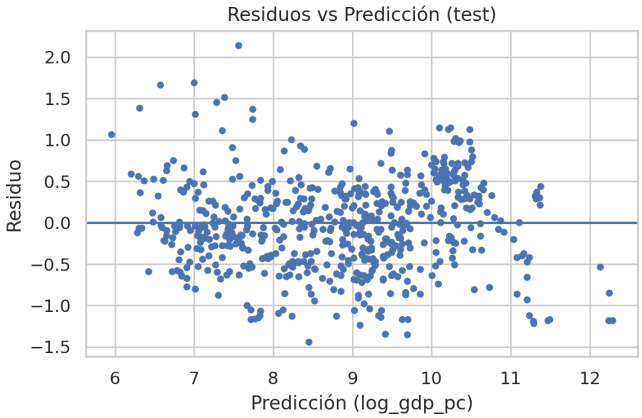
- $R^2 \approx 0{,}8494$
- MSE $\approx 0{,}2906$
- MAE $\approx 0{,}4244$

Diagnóstico gráfico

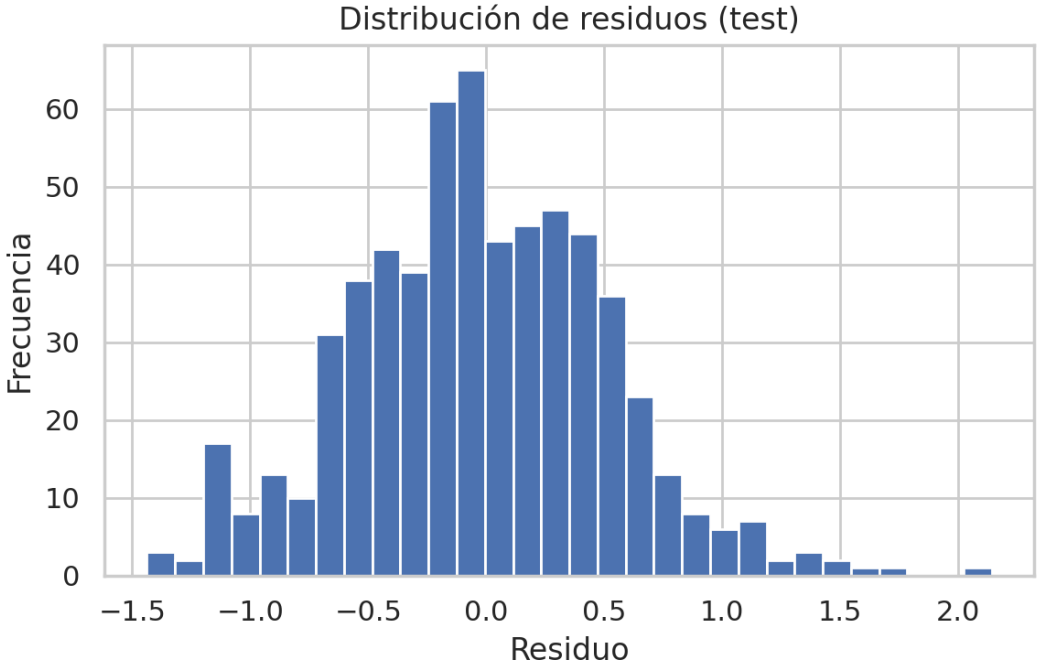
Se revisó la calidad predictiva (*predicho vs real*) y el comportamiento de los residuos. En general, el gráfico predicho vs real muestra buena alineación con la diagonal, mientras que los residuos se concentran alrededor de cero, con algunos valores extremos consistentes con outliers.



(a) Predicho vs real (test)



(b) Residuos vs predicción (test)



(c) Distribución de residuos (test)

Figura 4: Diagnóstico del modelo OLS.

Limitaciones y notas

- **Faltantes**: la regresión utiliza solo filas completas (2441), por lo que se descartan observaciones con NaN en explicativas.
- **Outliers**: existen valores extremos (por ejemplo, inflación muy alta en ciertos episodios) que pueden afectar sensibilidad.

- **Interpretación**: el análisis es asociativo; no identifica causalidad.
- **Multicolinealidad**: el *condition number* elevado sugiere posible multicolinealidad; puede evaluarse con VIF como extensión.

Reproducibilidad

El flujo es reproducible: la descarga se realiza desde la API del Banco Mundial, se construye el panel, se guardan CSV intermedios/finales y se exportan figuras y métricas del modelo en la carpeta `outputs/`. Esto permite replicar resultados y actualizar el panel si se extiende el rango de años o se agregan nuevos indicadores.