Asignatura	Datos del alumno	Fecha
Diplomado en Analítica de Datos en la Gestión Empresarial	Apellidos: López Fernández	20/10/2025
	Nombre: Douglas	

Fundación Universitaria Internacional de La Rioja

Pregrado en Ingeniería Informática

Diplomado - Analítica de Datos en la Gestión Empresarial

Actividad 3: Creando un Storytelling con Datos

Douglas López Fernández

Laura Marcela Barona Marmolejo

Deisy Johanna Villarraga Cuaycan

Oscar David Bocanegra Capera

GLENN ELMER HERNANDEZ CAMELO

2025

Asignatura	Datos del alumno	Fecha
Diplomado en Analítica de Datos en la Gestión Empresarial	Apellidos: López Fernández	20/10/2025
	Nombre: Douglas	

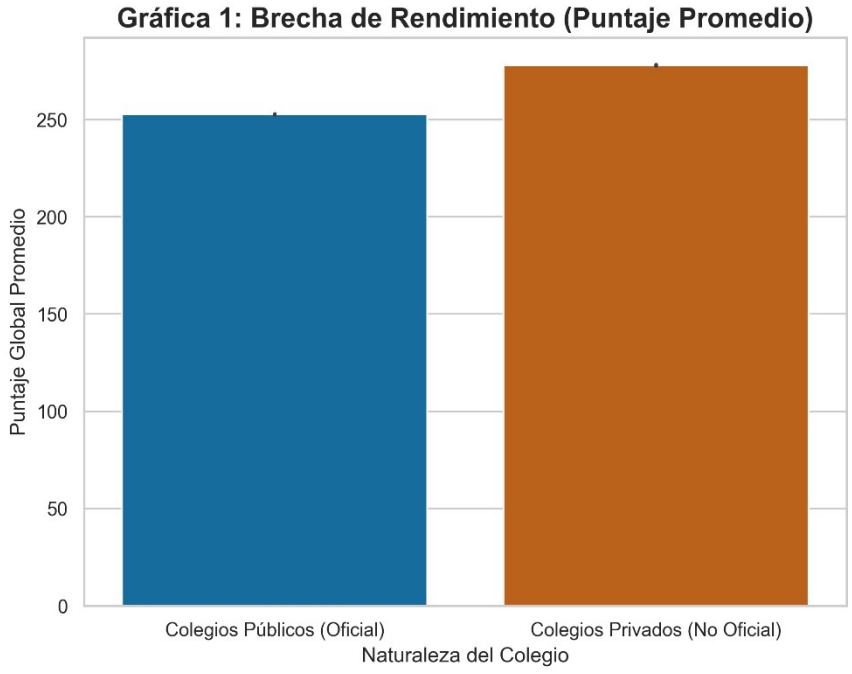
Introducción: El Punto de Partida No Es Igual para Todos

En Colombia, la educación es la promesa de equidad y movilidad social. Pero, ¿tienen todos nuestros jóvenes el mismo punto de partida? Este análisis busca responder una pregunta crucial: ¿Qué tan profunda es la brecha de rendimiento académico entre los estudiantes de colegios públicos (Oficiales) y privados (No Oficiales) en las principales ciudades del país? El objetivo no es solo mostrar números, sino cuantificar una inequidad estructural para informar a los gestores de políticas públicas.

Hallazgo 1: La Brecha es Evidente y Cuantificable

La primera fotografía de los datos es clara. Al comparar los puntajes globales promedio de las pruebas Saber 11, encontramos una diferencia significativa. Los colegios 'No Oficiales' (privados) obtienen, en promedio, **281** puntos, mientras que los 'Oficiales' (públicos) alcanzan **248** puntos. **Esto representa una brecha de rendimiento de 33 puntos,** una diferencia sustancial que marca el punto de partida de nuestro análisis.

Gráfica 1 (Barras)



Asignatura	Datos del alumno	Fecha
Diplomado en Analítica de Datos en la Gestión Empresarial	Apellidos: López Fernández	20/10/2025
	Nombre: Douglas	

Hallazgo 2: La Brecha No es Solo un Promedio, es una Distribución Completa

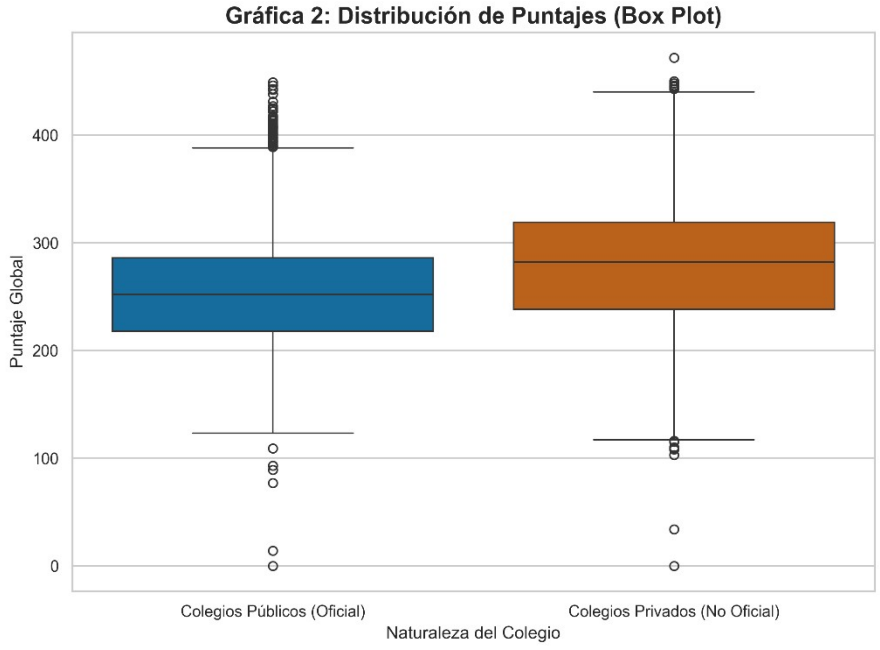
Un promedio puede ser engañoso. La verdadera historia está en la distribución de los resultados. Este diagrama de cajas (Box Plot) revela que la brecha es más profunda de lo que parece:

1. **Consistencia vs. Dispersión:** La 'caja' de los colegios privados es más compacta y está ubicada mucho más arriba, indicando resultados consistentemente altos. En contraste, la caja de los colegios públicos es más ancha y baja, mostrando una mayor dispersión y resultados menos predecibles.

2. **El 50% Central:** La mitad de los estudiantes de colegios privados (la caja completa) obtiene puntajes superiores a los que alcanza el 75% de los estudiantes de colegios públicos.

3. **La Mediana:** La línea central de la caja de los privados (la mediana) está casi al nivel del "bigote" superior de los públicos. Esto significa que un estudiante "promedio" de un colegio privado supera a la gran mayoría de los estudiantes del sector oficial.

Gráfica 2 (Boxplot)



Hallazgo 3: La Diferencia es Estadísticamente Real

La brecha que observamos en los gráficos es evidente, pero ¿podría ser producto del azar? Para confirmar nuestros hallazgos con rigor académico, realizamos una Prueba T de Student. El resultado (un p-valor extremadamente bajo, cercano a cero) nos permite afirmar con más del 99% de confianza que la diferencia de 33 puntos entre los promedios no es una casualidad. La brecha es, por lo tanto, estadísticamente significativa, lo que refuerza la necesidad de analizar sus causas estructurales.

Hallazgo 4: El Estrato Socioeconómico Agudiza la Brecha

Finalmente, exploramos si el estrato socioeconómico de las familias influye en esta brecha. El análisis es concluyente:

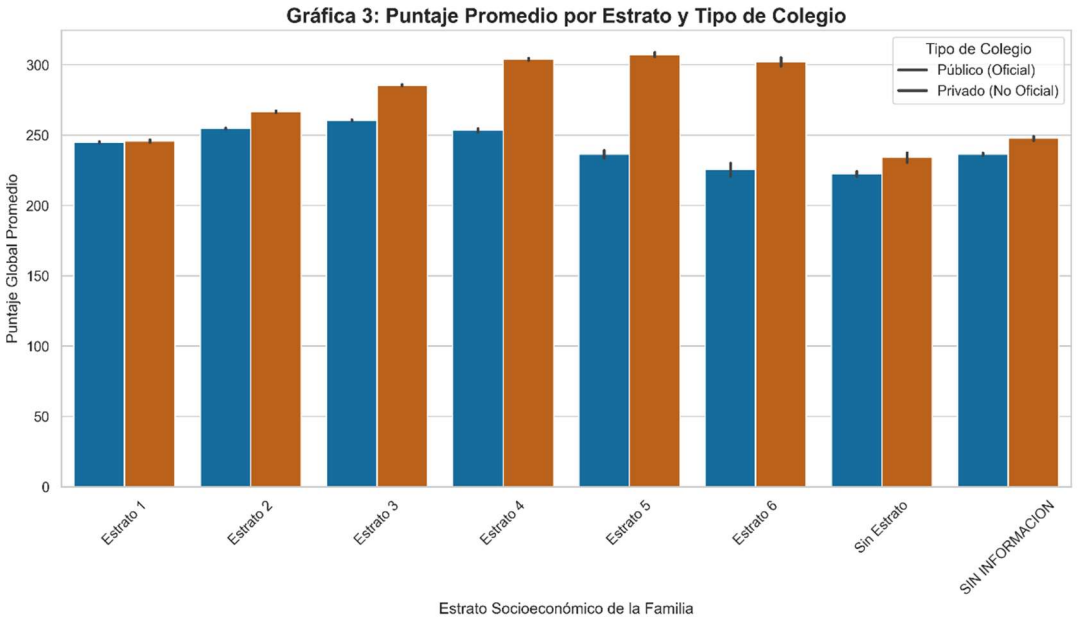
- La brecha de rendimiento entre colegios públicos y privados **existe en todos los estratos**.

- Sin embargo, la diferencia se acentúa a medida que el estrato es más bajo. Un estudiante de estrato 1 o 2 en un colegio privado tiene una ventaja de puntaje mucho mayor sobre su par en un colegio público, en comparación con la que se observa en los estratos 5 o 6.

Esto sugiere que, si bien la naturaleza del colegio es un factor clave, el contexto socioeconómico del estudiante actúa como un multiplicador de la inequidad.

Asignatura	Datos del alumno	Fecha
Diplomado en Analítica de Datos en la Gestión Empresarial	Apellidos: López Fernández	20/10/2025
	Nombre: Douglas	

Gráfica 3 (Estratos)



Conclusión y Llamado a la Acción

Nuestra historia con datos comenzó con una pregunta sobre equidad y termina con una evidencia clara y validada estadísticamente:

1. **Confirmamos** una brecha de rendimiento de **33 puntos** entre la educación pública y privada.

2. **Validamos** que esta diferencia **no es producto del azar**, sino un fenómeno estadísticamente real.

3. **Demostramos** que el estrato socioeconómico es un factor crucial que intensifica esta brecha.

El llamado a la acción para los gestores de políticas públicas es urgente. No basta con invertir en 'calidad'; se requiere una **estrategia focalizada de nivelación en la educación oficial**, con recursos priorizados en las zonas y estratos que demuestran mayor rezago. Los datos no mienten: la cancha no está pareja.

--- **RESULTADO ANÁLISIS LISTO** ---

Total, de registros para el análisis (año 20224): 272984

COLE_NATURALEZA

OFICIAL 173152

NO OFICIAL 99832

--- **RESULTADOS DE LA PRUEBA T-STUDENT** ---

Promedio Público (Oficial): 252.69

Promedio Privado (No Oficial): 277.89

Diferencia: 25.20 puntos

Valor P (p-value): 0.0

Conclusión Estadística: La diferencia ES estadísticamente significativa.

Anexo: Justificación Técnica de las Herramientas Utilizadas

Para el desarrollo de este análisis de datos, se seleccionó el lenguaje de programación **Python** junto con su ecosistema de bibliotecas especializadas. Esta elección no es arbitraria, sino que responde a estándares de la industria de la analítica de datos que garantizan la escalabilidad, reproducibilidad y robustez del estudio.

A continuación, se detallan las herramientas específicas y su rol en el proyecto:

Asignatura	Datos del alumno	Fecha
Diplomado en Analítica de Datos en la Gestión Empresarial	Apellidos: López Fernández	20/10/2025
	Nombre: Douglas	

1. **Python: El Lenguaje de Programación**

   Python fue elegido como el lenguaje base por ser de código abierto, de propósito general y por contar con una de las comunidades de desarrolladores y científicos de datos más grandes del mundo. Esto garantiza acceso a una vasta cantidad de documentación, soporte y bibliotecas de vanguardia para cualquier desafío analítico.

2. **Pandas: Manipulación y Limpieza de Datos**

   La biblioteca **Pandas** fue el pilar para el tratamiento de los datos. Su estructura de datos principal, el **DataFrame**, es una herramienta altamente optimizada para manejar datos tabulares. Las razones clave para su uso fueron:

   - **Manejo de Grandes Volúmenes de Datos:** El conjunto de datos original *("**Resultados_únicos_Saber_11_20251020.csv**")* se obtiene de la búsqueda de "***Resultados Saber 11***" la cual superaba los 3 GB, un tamaño que excede la capacidad de la memoria RAM de la mayoría de los computadores personales si se intenta cargar de una sola vez. Se utilizó la funcionalidad de procesamiento por trozos (chunks) de Pandas, que permite leer, filtrar y procesar el archivo por partes, asegurando que el análisis fuera posible sin requerir hardware especializado.

   - **Eficiencia en Limpieza y Filtrado:** Se utilizaron operaciones vectorizadas de Pandas para normalizar texto, filtrar las cinco ciudades principales y eliminar registros nulos de manera eficiente, tareas que serían extremadamente lentas y propensas a errores en herramientas tradicionales como hojas de cálculo.

3. **Matplotlib y Seaborn: Visualización de Datos**

   Para la generación de las visualizaciones, se utilizó una combinación de **Matplotlib** y **Seaborn**:

- **Matplotlib:** Es la biblioteca fundamental para la creación de gráficos en Python. Se utilizó como motor base para configurar los lienzos y ejes de las visualizaciones.

- **Seaborn:** Es una biblioteca de alto nivel construida sobre Matplotlib, especializada en la creación de gráficos estadísticos atractivos e informativos. Fue la herramienta principal para generar el gráfico de barras, el diagrama de cajas (Box Plot) y el gráfico de barras agrupado por estrato, ya que simplifica la sintaxis y produce visualizaciones estéticamente superiores y más fáciles de interpretar.

4. SciPy: Validación Estadística

Para elevar el rigor del análisis más allá de la simple descripción, se empleó la biblioteca SciPy. Específicamente, su módulo stats se utilizó para realizar la Prueba T de Student para muestras independientes. Esta prueba permitió validar con un alto nivel de confianza que la diferencia observada en los puntajes promedio entre colegios oficiales y no oficiales no era producto del azar, dotando de significancia estadística a las conclusiones del estudio.

Conclusión de la Justificación

El uso de este conjunto de herramientas (Python + Pandas, Seaborn, SciPy) no solo permitió ejecutar el análisis de manera eficiente y escalable, sino que también garantiza la reproducibilidad del estudio. Cualquier investigador con acceso a los mismos datos puede ejecutar el mismo script y obtener exactamente los mismos resultados y visualizaciones, un principio fundamental en la ciencia de datos.