Pontificia Universidad
JAVERIANA
Colombia

Reporte 1: Análisis de variables predictivas del rendimiento académico

Juan Pablo Arias Buitrago

Sergio Pardo Hurtado

Facultad de Ciencias Básicas

033520: Análisis de Regresión

Gabriel Camilo Pérez Castañeda, Phd

Pontificia Universidad Javeriana

Facultad de Ciencias

Bogotá D.C.

13 de septiembre de 2025

1. Análisis Exploratorio de Datos (EDA)

1.1. Análisis Inicial

A partir de la inspección inicial con la función `str(df)` aplicada sobre el conjunto de datos se obtiene la Tabla 1, donde se observa la estructura general del mismo, incluyendo el tipo de variable y algunos ejemplos de los valores que éste toma.

Variable	**Tipo**	**Descripción**
ID_Alumno	Caracter (chr)	Identificador único del estudiante ('S1000', 'S1100')
Edad	Numérico	Edad del estudiante en años (23, 20, 21, 18)
Genero	Caracter (chr)	Género del estudiante (Female, Male, Other)
Horas_Estudio	Numérico	Horas de estudio promedio por día (0, 6.9, 5)
Redes_Sociales	Numérico	Horas dedicadas a redes sociales por día (1.2, 3.1, 4.4)
Netflix	Numérico	Horas dedicadas a ver Netflix por día (1.1, 1.3, 0.5)
Trabajo	Caracter (chr)	Indica si el estudiante tiene trabajo de medio tiempo (Yes, No)
Asistencia	Numérico	Porcentaje de asistencia a clases (85, 97.3, 71)
Horas_Sueño	Numérico	Horas de sueño promedio por día (8, 4.6, 9.2)
Calidad_Dieta	Caracter (chr)	Calidad de la dieta del estudiante (Fair, Good, Poor)
Frecuencia_Ejercicio	Numérico	Frecuencia de ejercicio semanal (6, 1, 4, 3)
Educacion_Parental	Caracter (chr)	Nivel educativo de los padres (None, High School, Bachelor, Master)
Calidad_Internet	Caracter (chr)	Calidad del internet en el hogar (Poor, Average, Good)
Salud_Mental	Numérico	Autoevaluación de la salud mental del estudiante en una escala de 1 a 10
Act_Extraescolar	Caracter (chr)	Participación en actividades extracurriculares (Yes, No)
Puntaje_Examen	Numérico	Puntaje obtenido en el examen (56.2, 100, 34.3)

Tabla 1: Descripción de las variables del dataset (n = 1000, p = 16).

1.2. Matriz de Correlación

Utilizando la función `corrplot(matrix)` se obtiene la matriz de correlación entre las variables numéricas, lo que permite identificar posibles relaciones lineales con la variable objetivo (`exam_score`) y evaluar la presencia de colinealidad entre predictores (Figura 1).

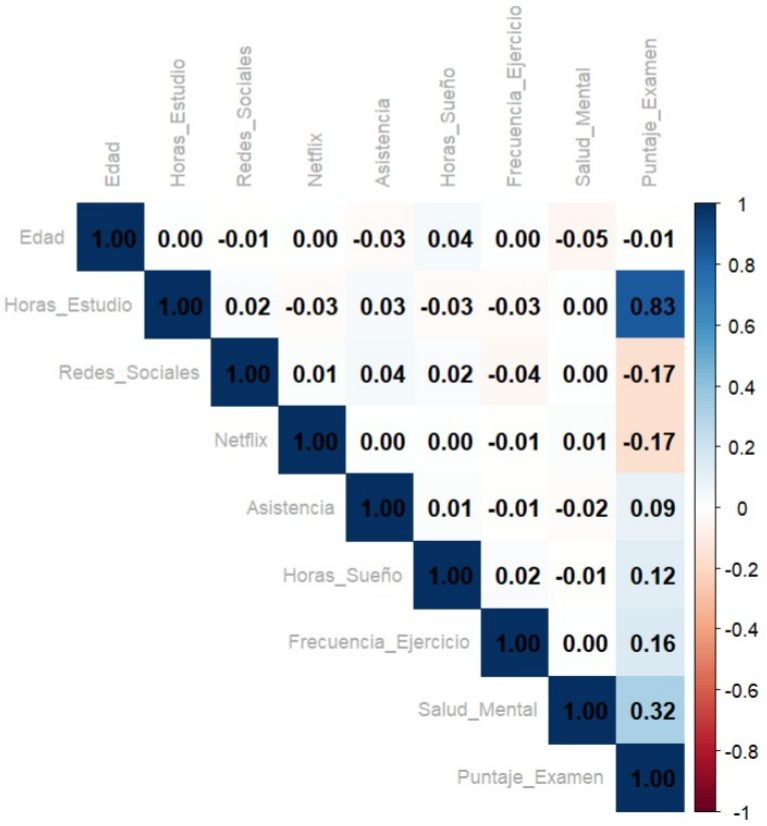

Figura 1: Matriz de Correlación

Observaciones:

- El puntaje obtenido en el examen muestra alta correlación con las horas de estudio por día ($r \approx 0{,}8$).
- Existe correlación moderada con la salud mental ($r \approx 0{,}32$) y correlaciones inversas leves con el uso de redes sociales y Netflix.
- Variables como asistencia a clase, horas de sueño, edad y ejercicio presentan correlaciones muy bajas con respecto al puntaje obtenido ($r < 0{,}2$).
- En general, se observa poca correlación entre las variables explicativas, lo que sugiere ausencia de multicolinealidad significativa.

1.3. Relación entre la Variable Objetivo y las Explicativas

Utilizando gráficos de dispersión (para variables numéricas) y gráficos de barras (para variables categóricas), se analiza la relación de cada variable explicativa con la variable objetivo `exam_score`.

1.3.1. Puntaje_Examen vs. age

De la Figura 2 se observa que no existe una relación lineal clara entre la edad de los estudiantes y el puntaje del examen, lo cual es coherente con la correlación obtenida entre ambas variables ($r \approx -0{,}01$). A priori, la edad parece no ser un factor relevante para explicar las variaciones en el desempeño académico.

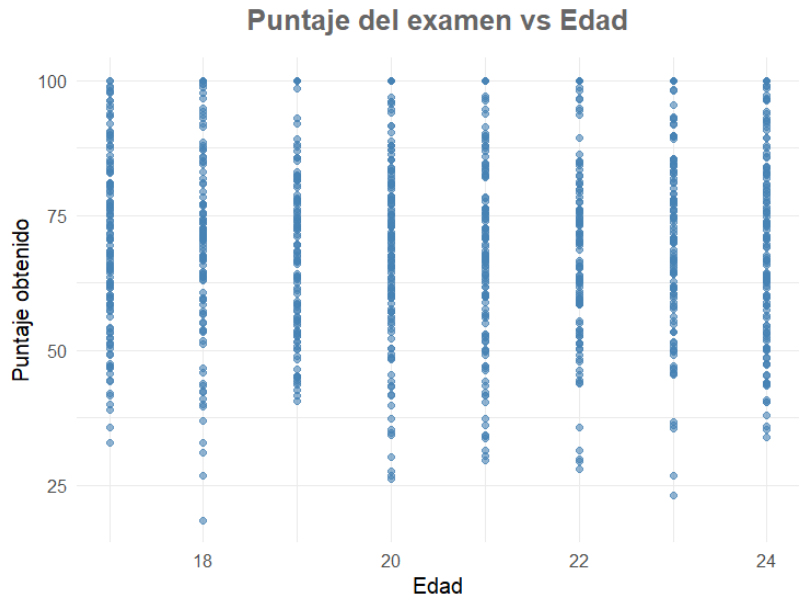

Figura 2: Relación entre `Puntaje_Examen` y `Edad`

1.3.2. Puntaje_Examen vs. Horas_Estudio

la Figura 3 evidencia una relación lineal positiva entre las horas de estudio por día y el puntaje obtenido. A medida que aumentan las horas de estudio, también se incrementa el puntaje, lo cual es consistente con la alta correlación estimada entre ambas variables ($r \approx 0{,}83$). Este resultado sugiere que el tiempo de estudio es una variable importante para explicar el desempeño académico.

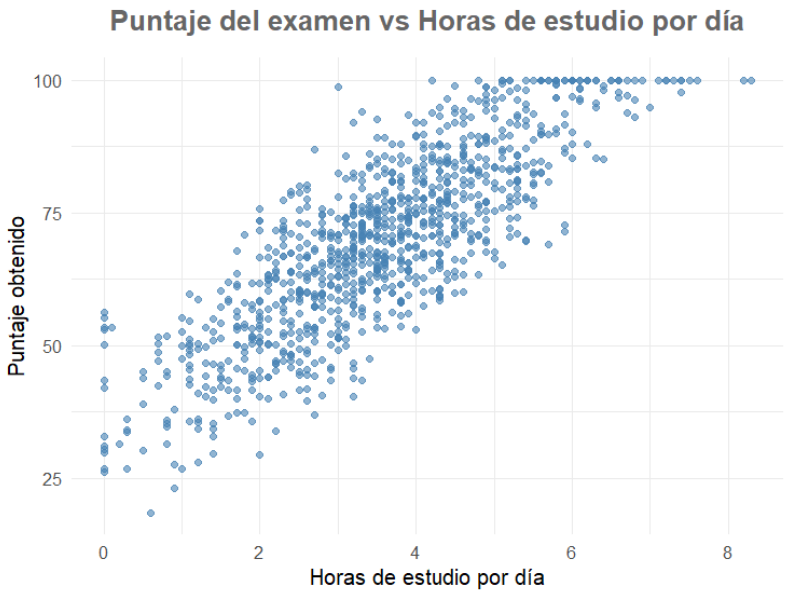

Figura 3: Relación entre `Puntaje_Examen` y `Horas_Estudio`

1.3.3. Puntaje_Examen vs. Redes_Sociales

En la Figura 4 no se observa una relación lineal clara entre las horas dedicadas a redes sociales y el puntaje obtenido, lo cual es consistente con la correlación estimada ($r \approx -0{,}17$). En consecuencia, las horas en redes sociales, de manera aislada, no parecen explicar el desempeño académico.

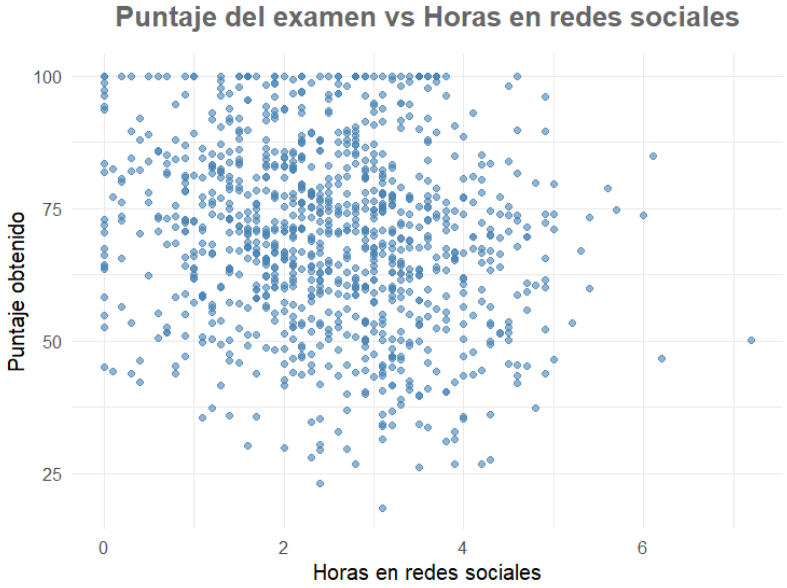

Figura 4: Relación entre `Puntaje_Examen` y `Redes_Sociales`

1.3.4. Puntaje_Examen vs. Netflix

De la Figura 5 se observa un patrón similar al encontrado con las horas en redes sociales: no se evidencia una relación lineal clara entre las horas dedicadas a ver Netflix y el puntaje obtenido en el examen. La correlación estimada, igualmente baja y negativa ($r \approx -0{,}17$), sugiere que, por sí sola, esta variable no explica de forma significativa el desempeño académico. Sin embargo, resulta pertinente considerar posibles transformaciones de la variable o explorar interacciones conjuntas entre el tiempo en Netflix y las horas en redes sociales, lo cual podría revelar relaciones más evidentes con respecto al puntaje.

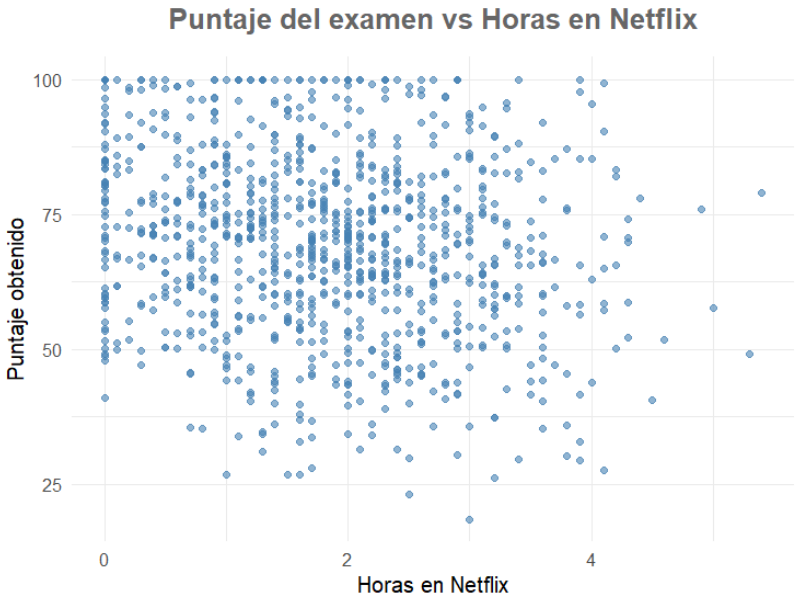

Figura 5: Relación entre `Puntaje_Examen` y `Netflix`

1.3.5. Puntaje_Examen vs. Asistencia

De la Figura 6 se observa una amplia dispersión de los datos y ausencia de una relación lineal evidente entre el porcentaje de asistencia a clases y el puntaje obtenido en el examen. La correlación estimada ($r \approx 0{,}09$) sugiere que esta variable presenta una capacidad explicativa limitada. Sin embargo, tanto desde la práctica como a partir del propio gráfico, se reconoce que una mayor asistencia tiende a asociarse con un mejor rendimiento académico.

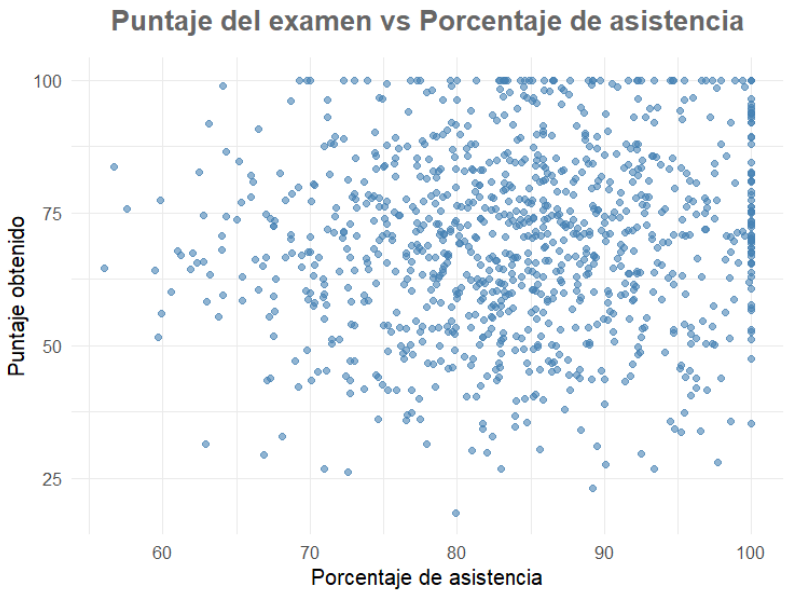

Figura 6: Relación entre `Puntaje_Examen` y `Asistencia`

1.3.6. Puntaje_Examen vs. Horas_Sueño

De la Figura 7 se puede interpretar que no existe una relación lineal clara entre las horas de sueño y el puntaje del examen. Aunque se observa cierta concentración de estudiantes con puntajes altos (superiores a 75) en un rango de 6 a 8 horas de sueño, también hay casos dispersos con puntajes bajos en ese mismo rango.

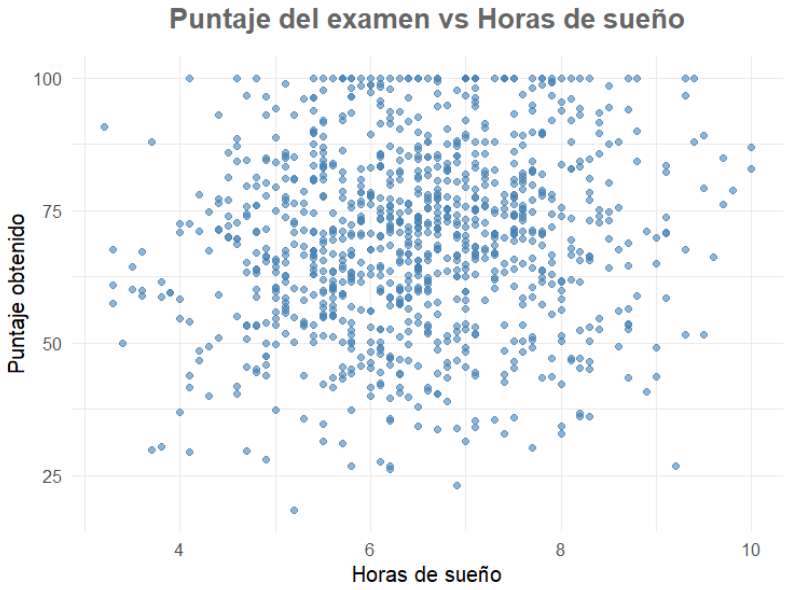

Figura 7: Relación entre `Puntaje_Examen` y `Horas_Sueño`

1.3.7. Puntaje_Examen vs. Frecuencia_Ejercicio

De la Figura 8 se puede interpretar no hay relación lineal entre las dos variables, en todos los niveles aparecen estudiantes con calificaciones altas y bajas, por lo que la frecuencia de ejercicio por sí sola no explica el rendimiento académico.

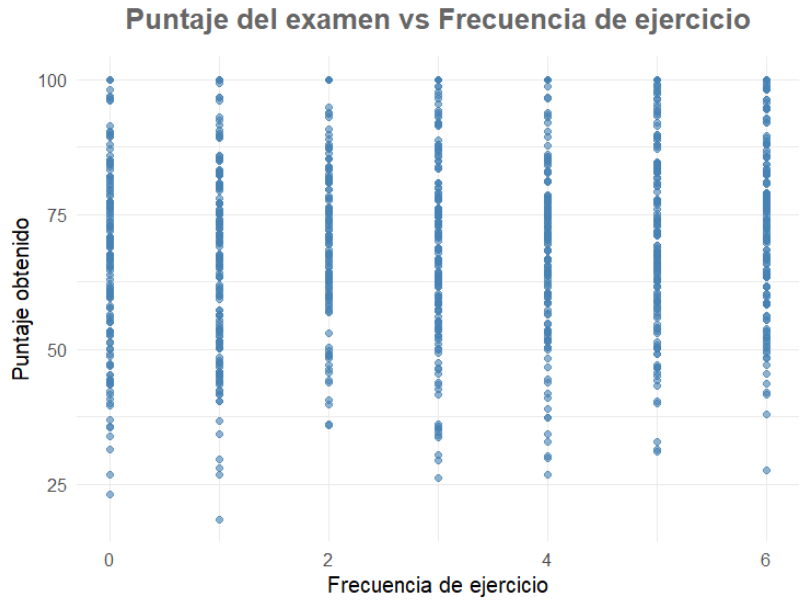

Figura 8: Relación entre `Puntaje_Examen` y `Frecuencia_Ejercicio`

1.3.8. Puntaje_Examen vs. Salud_Mental

De la misma manera, para la figura 9 tenemos que no hay relación lineal entre las variables de puntaje y salud mental, de igual forma podemos ver que en la variable de salud mental aparecen estudiantes con diferentes niveles de clasificación, por lo que el nivel de ejercicio por sí solo no explica el rendimiento académico.

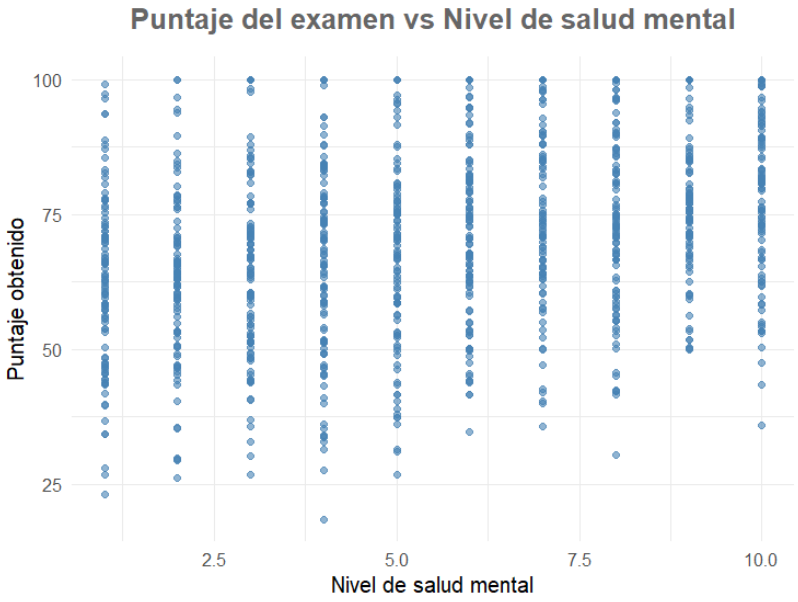

Figura 9: Relación entre `Puntaje_Examen` y `Salud_Mental`

1.3.9. Puntaje_Examen vs. Género

En la figura 10 se puede observar que el promedio del puntaje no varía mucho respecto a los géneros establecidos en el dataset.

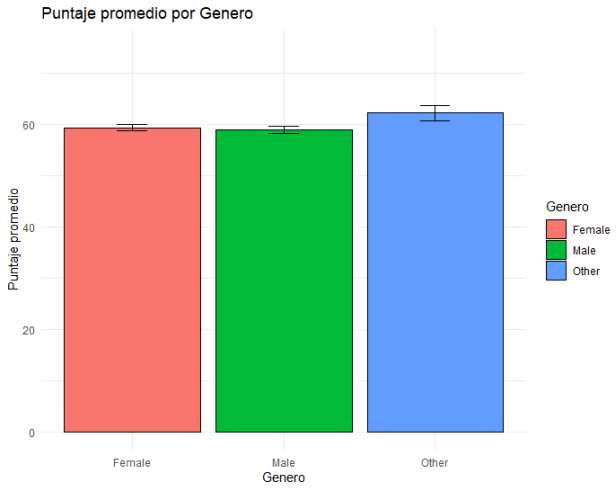

Figura 10: Relación entre `Puntaje_Examen` y `Género`

1.3.10. Puntaje_Examen vs. Trabajo

En la figura 11 se puede observar que, en este caso, el hecho de que el estudiante tenga un trabajo de medio tiempo o no, no varía demasiado el puntaje promedio.

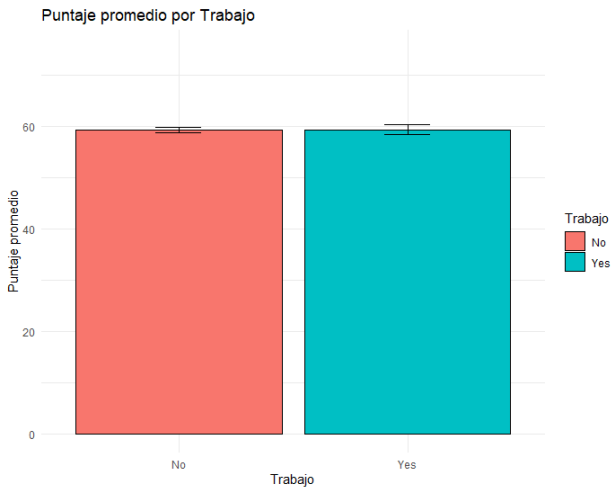

Figura 11: Relación entre `Puntaje_Examen` y `Trabajo`

1.3.11. Puntaje_Examen vs. Calidad_Dieta

A partir de la figura 12, podemos observar que la categoría de dieta *"Fair"* se encuentra levemente por encima de los otros dos tipos de dieta en cuanto al puntaje promedio.

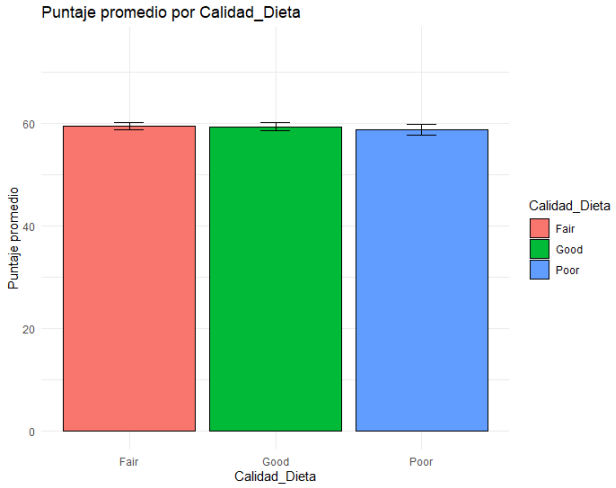

Figura 12: Relación entre `Puntaje_Examen` y `Calidad_Dieta`

1.3.12. Puntaje_Examen vs. Calidad_Internet

En la figura 13 se pueden observar las categorías de la variable `Calidad_Internet`. Además, podemos notar que la categoría *"Poor"* tiene un puntaje promedio levemente superior al de las demás.

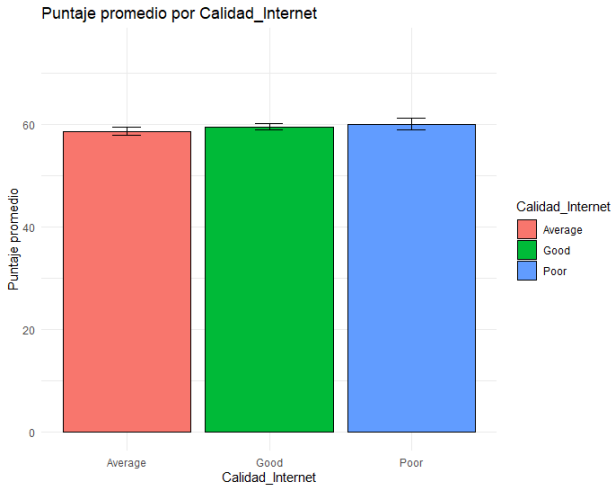

Figura 13: Relación entre `Puntaje_Examen` y `Calidad_Internet`

1.3.13. Puntaje_Examen vs. Actividades_Extraescolares

A partir de la figura 14, podemos observar que, aunque ambos grupos presentan promedios similares, el puntaje promedio es más alto en los estudiantes que no realizan ninguna actividad extracurricular.

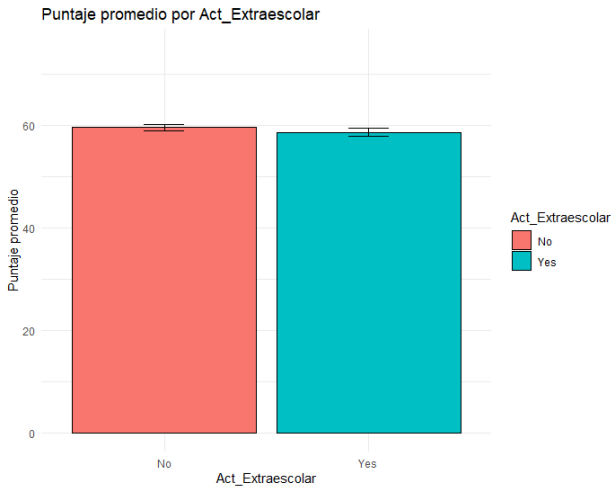

Figura 14: Relación entre `Puntaje_Examen` y `Actividades_Extraescolares`

1.3.14. Puntaje_Examen vs. Nivel_Educacion_Parental

En la figura 15 se pueden observar las cuatro categorías presentes en la variable `Nivel_Educacion_Parental`: *"Bachelor"*, *"High School"*, *"Master"* y *"None"*. Al contrastarlas con el puntaje promedio del examen, no se evidencian diferencias significativas entre las categorías. La categoría con el puntaje promedio más bajo es *"High School"*, seguida de *"Bachelor"*.

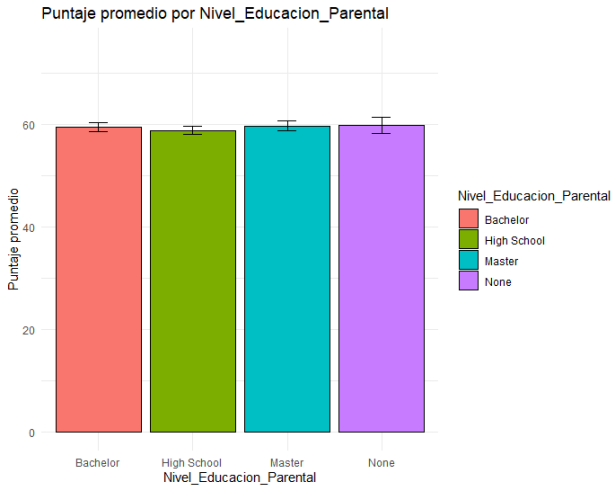

Figura 15: Relación entre `Puntaje_Examen` y `Nivel_Educacion_Parental`

1.4. Transformación de Variables

Con base en el análisis exploratorio de los datos, se sugieren las siguientes transformaciones y combinaciones de variables para capturar mejor las posibles relaciones no lineales y las interacciones que podrían estar influyendo en el `Puntaje_Examen`.

- **Horas de Estudio (`Horas_Estudio`)**: Presenta una fuerte correlación positiva con el puntaje ($r \approx 0{,}83$). Dado que el efecto ya es claro y consistente, no se considera necesario aplicar transformaciones en esta variable por el momento.

- **Redes Sociales y Netflix (`Redes_Sociales`, `Netflix`)**: Individualmente muestran correlaciones bajas ($r \approx -0{,}17$). Se sugiere combinarlas en una nueva variable:

$$Tiempo\_Pantalla = Redes\_Sociales + Netflix$$

  con el fin de capturar el efecto global del tiempo de ocio frente a pantallas.

- **Asistencia (`Asistencia`)**: La correlación con el puntaje es débil ($r \approx 0{,}09$). No obstante, en la práctica una mayor asistencia tiende a asociarse con un mejor rendimiento.Se aplicara una transformación cuadrática ($Asistencia^2$) para capturar efectos no lineales.

- **Horas de Sueño (`Horas_Sueño`)**: El análisis hecho del gráfico de dispersión sugieren una relación en forma de U invertida (óptimo entre 6–8 horas).

$$Horas\_\text{Sueño}^2$$

- **Ejercicio y Salud Mental (`Frecuencia_Ejercicio`, `Salud_Mental`)**: Aunque individualmente presentan correlaciones bajas, podrían tener un efecto combinado en el desempeño.

$$Frecuencia\_Ejercicio \times Salud\_Mental$$

- **Variables categóricas (`Genero`, `Calidad_Dieta`, `Educacion_Parental`, `Calidad_Internet`, `Act_Extraescolar`, `Trabajo`)**: Se transformarán en variables *dummy* mediante codificación `one-hot`, generando $k-1$ columnas adicionales por cada categoría. Por ejemplo, si `Calidad_Dieta` presenta las categorías *fair*, *Good* y *Poor*, se crearán:

$$Dieta\_fair, \quad Dieta\_Good$$

tomando a *Poor* como categoría de referencia.

1.4.1. Modelo propuesto con transformaciones

Considerando las transformaciones sugeridas, el modelo de regresión extendido se expresa en la Tabla 2.

Componente	**Término en el modelo**
Intercepto	β_0
Horas de Estudio	$\beta_1 \cdot Horas\_Estudio$
Horas de Estudio (cuadrática)	$\beta_2 \cdot Horas\_Estudio^2$
Tiempo de Pantalla (Redes + Netflix)	$\beta_3 \cdot (Redes\_Sociales + Netflix)$
Asistencia	$\beta_4 \cdot Asistencia$
Asistencia (cuadrática)	$\beta_5 \cdot Asistencia^2$
Horas de Sueño	$\beta_6 \cdot Horas_$Sueño
Horas de Sueño (cuadrática)	$\beta_7 \cdot Horas_$Sueño^2
Interacción Ejercicio × Salud Mental	$\beta_8 \cdot (Frecuencia\_Ejercicio \times Salud\_Mental)$
Género	$\sum_j \gamma_j \cdot Genero_j$
Calidad de la Dieta	$\sum_k \delta_k \cdot Dieta_k$
Educación Parental	$\sum_l \theta_l \cdot Educacion\_Parental_l$
Calidad del Internet	$\sum_m \phi_m \cdot Calidad\_Internet_m$
Actividad Extraescolar	$\sum_n \psi_n \cdot Act\_Extraescolar_n$
Trabajo	$\sum_o \omega_o \cdot Trabajo_o$
Error	$+\varepsilon$

Tabla 2: Ecuación del modelo con transformaciones propuestas

2. Estimación de modelos, ajuste y validación

2.1. Estimación del Modelo Completo

Para evaluar la capacidad explicativa del modelo, se definieron dos especificaciones: (i) un **Mod_Completo** que incluye únicamente las variables originales, y (ii) un **Mod_Comp_Transformado**, que incorpora transformaciones no lineales (cuadráticas), la variable compuesta *Tiempo_Pantalla* y la interacción *Frecuencia_Ejercicio* × *Salud_Mental*.

La comparación entre ambos modelos se realizó empleando los criterios estadísticos: **R^2**, **R^2 ajustado**, **RMSE** (raíz del error cuadrático medio) y el **AIC** (Criterio de Información de Akaike). Los resultados obtenidos se presentan en la Tabla 3.

Modelo	**AIC**	**R^2**	**R^2 ajustado**	**RMSE**
Mod_Completo	6210.94	0.902	0.900	5.288
Mod_Comp_Transformado	6180.18	0.905	0.903	5.192

Tabla 3: Comparación de desempeño entre Mod_Completo y Mod_Comp_Transformado.

Se observa que el **Mod_Comp_Transformado** presenta un *AIC* menor, así como ligeras mejoras en el R^2 *ajustado* y el *RMSE*, lo que indica un mejor ajuste sin evidencia de sobreparametrización. Por lo tanto, con base en estos criterios de comparación, se selecciona el **Mod_Comp_Transformado** como el primer modelo de análisis principal, a partir del cual se evaluará la significancia global e individual de los parámetros.

2.1.1. Significancia Global del Modelo

La significancia global se evaluó mediante un *F-test* (ANOVA) aplicado al modelo extendido. El resultado muestra un estadístico $F = 425$ con 22 y 977 grados de libertad y un valor-$p < 2{,}2 \times 10^{-16}$, lo que indica que, en conjunto, las variables explicativas aportan información significativa para explicar la variabilidad del puntaje en el examen.

2.1.2. Significancia Individual de Parámetros

En cuanto a la significancia individual, se analizaron los coeficientes estimados mediante *t-tests*. Se observa que las variables más relevantes y altamente significativas (valor-$p < 0{,}001$) incluyen:

- Horas de estudio ($\beta = 11{,}81$, $p < 2 \times 10^{-16}$).
- Cuadrado de horas de estudio ($\beta = -0{,}31$, $p < 0{,}001$), lo que evidencia un efecto no lineal.
- Tiempo en pantalla ($\beta = -2{,}45$, $p < 2 \times 10^{-16}$).
- Frecuencia de ejercicio ($\beta = 1{,}91$, $p < 2 \times 10^{-16}$).
- Salud mental ($\beta = 2{,}19$, $p < 2 \times 10^{-16}$).
- Interacción entre ejercicio y salud mental ($\beta = -0{,}08$, $p < 0{,}01$).

En contraste, otras variables como género, asistencia, educación parental y actividad extracurricular no mostraron evidencia estadística suficiente para considerarse significativas en el modelo.

Estos hallazgos refuerzan que las horas de estudio, el tiempo de pantalla y la interacción entre hábitos de salud (ejercicio y salud mental) son los predictores más influyentes en el desempeño académico.

2.2. Eliminación de Variables No Significativas

Con el fin de simplificar el modelo y mejorar su interpretabilidad, se aplicó el procedimiento `stepAIC` para realizar una selección automática de variables en ambas direcciones (*forward* y *backward*). Este proceso eliminó aquellas variables que no aportaban significancia estadística al modelo, manteniendo únicamente los predictores con mayor contribución a la explicación de la variable respuesta.

En la Tabla 4 se comparan los tres modelos: el **Modelo Completo**, el **Modelo Transformado Completo** y el **Modelo Transformado Reducido**. Se utilizaron como criterios de evaluación el *AIC*, el R^2, el R^2 *ajustado* y el *RMSE*.

Modelo	**AIC**	**R^2**	**R^2 Ajustado**	**RMSE**
Mod_Completo	6210.938	0.9018	0.8999	5.288
Mod_Comp_Transformado	6180.178	0.9054	0.9033	5.192
Mod_Transformado_Reducido	6161.542	0.9045	0.9037	5.217

Tabla 4: Comparación de desempeño entre modelos completos y reducido.

Se observa que el **Modelo Transformado Reducido** presenta el menor *AIC*, lo cual indica un mejor balance entre ajuste y complejidad. Asimismo, el R^2 *ajustado* se mantiene prácticamente igual respecto al modelo transformado completo, mientras que el *RMSE* varía de forma mínima. Esto confirma que la eliminación de variables no significativas no afecta de manera sustancial la capacidad predictiva del modelo.

2.3. Validación de Supuestos del Modelo

En el modelo se aplicaron distintas pruebas para evaluar los supuestos de la regresión:

- La normalidad de los residuos se verificó con Shapiro-Wilk, Anderson-Darling y Kolmogorov-Smirnov: en todos ellos la hipótesis nula (H_0) supone que los residuos siguen una distribución normal, mientras que la alternativa (H_1) plantea que no lo hacen; se rechaza H_0 cuando el valor p es menor a 0.05.

- La homocedasticidad se contrastó con las pruebas de Breusch-Pagan y de varianza no constante: aquí H_0 indica varianza constante de los errores y H_1 la presencia de heterocedasticidad, rechazándose H_0 con valores p bajos.

- La independencia de errores se evaluó mediante la prueba de Durbin-Watson y la función de autocorrelación: en este caso H_0 establece ausencia de autocorrelación y H_1 presencia de correlación serial; valores p pequeños permiten rechazar H_0.

- La multicolinealidad se midió con el Factor de Inflación de la Varianza (VIF), el cual corresponde a un indicador que mide cuánto aumenta la varianza de los coeficientes por la correlación entre regresores; valores cercanos a 1 indican ausencia de colinealidad, mientras que valores superiores a 10 se consideran problemáticos.

- Finalmente, la influencia de observaciones individuales se analizó con la distancia de Cook y el gráfico de influencia, identificando como puntos influyentes aquellos que superan el umbral de $4/n$, ya que pueden distorsionar significativamente el ajuste del modelo.

2.3.1. Resultados

- **Normalidad de los residuos:** Las pruebas de Shapiro-Wilk ($p = 0{,}5103$), Anderson-Darling ($p = 0{,}5566$) y Kolmogorov-Smirnov ($p = 0{,}8384$) no permiten rechazar la hipótesis nula de normalidad. El gráfico Q-Q (Figura 16) también respalda esta conclusión. Sin embargo, dado que estas pruebas pueden ser sensibles al tamaño de muestra, se requiere un análisis más detallado.

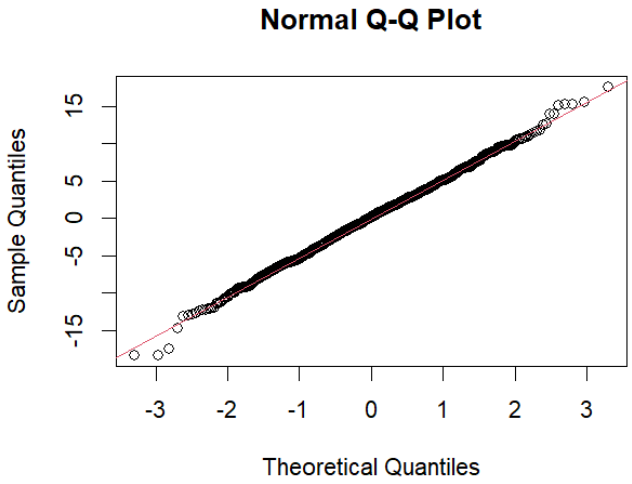

Figura 16: Gráfico Q-Q de los residuos del modelo reducido.

- **Homocedasticidad:** El test de Breusch-Pagan ($p = 0{,}032$) sugiere la presencia de heterocedasticidad, mientras que el test de varianza no constante ($p = 0{,}909$) no reporta problemas. La contradicción entre pruebas implica que la homogeneidad de varianza no está plenamente garantizada. Esto constituye una limitación relevante, pues la heterocedasticidad afecta directamente la eficiencia de los estimadores y la validez de los intervalos de confianza, por lo cual sería necesario aplicar métodos robustos o transformaciones adicionales en futuros análisis.

- **Independencia de los errores:** El test de Durbin-Watson ($DW = 1{,}97$, $p = 0{,}355$) indica ausencia de autocorrelación en los residuos, lo cual es consistente con la inspección de la función de autocorrelación. Este supuesto se considera cumplido.

- **Multicolinealidad:** Los valores de VIF resultaron moderadamente altos en *Horas de Estudio* y su término cuadrático ($VIF \approx 13$). Esto es esperable debido a la colinealidad inducida entre una variable y su transformación polinómica, pero plantea una limitación en la interpretación precisa de los coeficientes. Se debe considerar cuidadosamente este aspecto al explicar los resultados, pues podría sesgar la identificación del efecto real de dichas variables.

- **Observaciones influyentes:** El gráfico de influencia (Figura 17) evidencia algunos

casos con valores elevados de leverage y distancia de Cook. Aunque no comprometen gravemente el ajuste global, estas observaciones influyentes deben analizarse individualmente para descartar la existencia de sesgos o patrones anómalos en los datos.

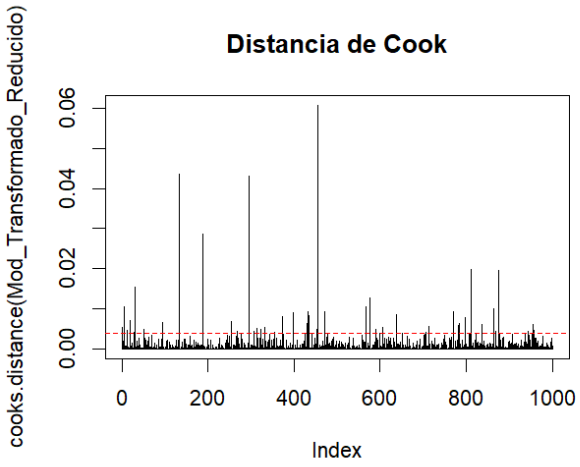

Figura 17: Identificación de observaciones influyentes en el modelo reducido.

En síntesis, aunque el modelo presenta un ajuste adecuado en términos generales, los resultados sugieren la necesidad de un análisis más riguroso en dos frentes críticos: la posible heterocedasticidad y la colinealidad entre variables transformadas. Estas limitaciones deben ser consideradas al momento de interpretar los coeficientes y al proponer extensiones o simplificaciones del modelo, de manera que se evite una sobre-interpretación de los resultados.

2.4. Elección del Modelo final e Interpretación de los Parámetros

Finalmente, tras aplicar las pruebas de reducción y validación de supuestos, se seleccionó como modelo final el **Modelo Transformado Reducido**. La interpretación de los parámetros estimados se presenta a continuación:

- **Intercepto** ($\hat{\beta}_0 = 0{,}63$**,** $p = 0{,}761$): No resulta estadísticamente significativo, por lo que carece de relevancia práctica en la explicación del modelo.
- **Horas de Estudio** ($\hat{\beta}_1 = 11{,}75$**,** $p < 0{,}001$): Cada hora adicional de estudio se asocia, en promedio, con un incremento de 11.75 puntos en el puntaje del examen, aunque este efecto se ve modificado por el término cuadrático.

- **Horas de Estudio al cuadrado** ($\hat{\beta}_2 = -0{,}30$, $p < 0{,}001$): El efecto marginal de estudiar más horas decrece conforme aumenta la cantidad de estudio. Esto indica la existencia de *rendimientos decrecientes*: a partir de cierto umbral, más horas de estudio aportan beneficios cada vez menores.

- **Tiempo en Pantalla** ($\hat{\beta}_3 = -2{,}46$, $p < 0{,}001$): Cada hora adicional frente a dispositivos electrónicos reduce, en promedio, 2.46 puntos en el puntaje del examen, evidenciando un efecto negativo y significativo.

- **Asistencia a clase** ($\hat{\beta}_4 = 0{,}15$, $p < 0{,}001$): Cada punto porcentual adicional en la asistencia está asociado con un aumento de 0.15 puntos en el examen. Su efecto acumulado resulta relevante en estudiantes con alta asistencia.

- **Horas de Sueño** ($\hat{\beta}_5 = 2{,}03$, $p < 0{,}001$): Cada hora adicional de sueño contribuye, en promedio, con un incremento de 2.03 puntos, mostrando la importancia del descanso en el rendimiento académico.

- **Frecuencia de Ejercicio** ($\hat{\beta}_6 = 1{,}93$, $p < 0{,}001$): A mayor frecuencia de actividad física, el puntaje del examen aumenta, con un efecto positivo y estadísticamente significativo.

- **Salud Mental** ($\hat{\beta}_7 = 2{,}21$, $p < 0{,}001$): Mejores niveles de salud mental se asocian con un incremento de 2.21 puntos en el examen, confirmando su impacto positivo.

- **Interacción Ejercicio–Salud Mental** ($\hat{\beta}_8 = -0{,}08$, $p = 0{,}0036$): El coeficiente negativo indica que, aunque ambos factores son positivos de manera individual, su efecto conjunto no es completamente aditivo. Es decir, realizar mucho ejercicio con un alto nivel de salud mental no aumenta el puntaje de manera lineal, sino que presenta un efecto atenuado.

En conjunto, el modelo explica un alto porcentaje de la variabilidad del puntaje de examen ($R^2_{ajustado} = 0{,}9037$), lo que sugiere que es una especificación robusta para capturar la relación entre las variables explicativas y el desempeño académico.

2.5. Conclusiones Iniciales

- **Horas de estudio**: constituyen el predictor más relevante, aunque con rendimientos decrecientes al aumentar en exceso (efecto cuadrático negativo).
- **Tiempo en pantalla**: presenta un impacto negativo significativo sobre el puntaje.
- **Horas de sueño**, **frecuencia de ejercicio** y **salud mental**: influyen positivamente en el desempeño.
- **Interacción ejercicio–salud mental**: indica la necesidad de equilibrio, pues niveles muy altos en ambos reducen el efecto positivo esperado.
- Variables como **género**, **educación parental**, **calidad del internet** y **actividades extracurriculares** no fueron estadísticamente significativas.
- La inclusión de **transformaciones e interacciones** mejoró el ajuste del modelo, reflejándose en menor AIC, mayor R^2 ajustado y menor RMSE.
- El **modelo reducido** resultó más parsimonioso y adecuado para explicar la relación entre los predictores y el *Puntaje_Examen*.

2.6. Origen del Dataset

El dataset **"Student Habits vs Academic Performance: A Simulated Study"** contiene información de 1,000 estudiantes. Cada registro representa a un alumno con variables relacionadas con sus hábitos diarios (horas de estudio, sueño, uso de redes sociales, calidad de la dieta, estado de salud mental, entre otros) y su puntaje final en el examen. Es un conjunto de datos limpio y estructurado, diseñado para análisis estadístico, exploración, visualización y modelado predictivo en proyectos de ciencia de datos. Resulta especialmente útil para estudios sobre cómo los hábitos de vida influyen en el rendimiento académico.

El diccionario de datos se encuentra tanto al principio del documento, en la sección de **Análisis Inicial**, como en el repositorio utilizado para el desarrollo de este proyecto, específicamente en el apartado **data**, bajo el nombre `data_dictionary.xlsx`.

2.6.1. Acceso al dataset en Kaggle

https://www.kaggle.com/datasets/student-habits-vs-academic-performance

2.6.2. Código Utilizado - Repositorio

El código empleado para generar los reportes, gráficas y modelos se encuentra en el siguiente repositorio: https://github.com/JuanParias29/Academic_Performance_Regression