



Análisis Exploratorio de Datos sobre: Student Performance Data



Autor:

Gabriel Alexander Matos Rincón



LinkedIn:

[linkedin.com/in/gabriel-matos-rincon13](https://www.linkedin.com/in/gabriel-matos-rincon13)



Fecha:

Julio, 2025



Descripción del Proyecto

Este proyecto tiene como finalidad aplicar los pasos fundamentales del **Análisis Exploratorio de Datos (EDA)** utilizando un conjunto de datos reales relacionados con el rendimiento académico de estudiantes. Se exploran variables como el promedio académico (GPA), número de ausencias, horas de estudio semanal, apoyo parental, tutoría y participación en actividades extracurriculares, entre otras.

El análisis se basa en el dataset **Student Performance Data**, disponible en [Kaggle](#), y está orientado a extraer patrones relevantes que puedan servir de base para modelos predictivos o para la toma de decisiones en contextos educativos.


1) Importando librerías necesarias

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Configuración Visual
sns.set(style="whitegrid")
pd.set_option('display.max_columns', None)
```

2) Importando el archivo


```
from google.colab import files
uploaded = files.upload()
```

 Elegir archivos



Ningún archivo seleccionado Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

Saving Student performance data.csv to Student performance data (1).csv

```
df = pd.read_csv(next(iter(uploaded)))
df.head()
```



	StudentID	Age	Gender	Ethnicity	ParentalEducation	StudyTimeWeekly	Absences	Tutoring	ParentalSupport	Extracurricular	Sports	Music	Volunteering	GPA	GradeClass
0	1001	17	1	0	2	19.833723	7	1	2	0	0	1	0	2.929196	2.0
1	1002	18	0	0	1	15.408756	0	0	1	0	0	0	0	3.042915	1.0
2	1003	15	0	2	3	4.210570	26	0	2	0	0	0	0	0.112602	4.0
3	1004	17	1	0	3	10.028829	14	0	3	1	0	0	0	2.054218	3.0
4	1005	17	1	0	2	4.672495	17	1	3	0	0	0	0	1.288061	4.0



3) Información general del dataset

```
print("Información general del dataset:")
print(df.info())

print("\nEstadísticas descriptivas:")
print(df.describe(include='all'))
```



Información general del dataset:

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2392 entries, 0 to 2391
Data columns (total 15 columns):
Column Non-Null Count Dtype
--- -
0 StudentID 2392 non-null int64
1 Age 2392 non-null int64
2 Gender 2392 non-null int64
3 Ethnicity 2392 non-null int64
4 ParentalEducation 2392 non-null int64
5 StudyTimeWeekly 2392 non-null float64
6 Absences 2392 non-null int64
7 Tutoring 2392 non-null int64
8 ParentalSupport 2392 non-null int64
9 Extracurricular 2392 non-null int64
10 Sports 2392 non-null int64
11 Music 2392 non-null int64
12 Volunteering 2392 non-null int64
13 GPA 2392 non-null float64
14 GradeClass 2392 non-null float64
dtypes: float64(3), int64(12)
memory usage: 280.4 KB
None

Estadísticas descriptivas:

StudentID

Age

Gender

Ethnicity

ParentalEducation

\

count

2392.000000

2392.000000

2392.000000

2392.000000

2392.000000

mean

2196.500000

16.468645

0.510870

0.877508

1.746237

std

690.655244

1.123798

0.499986

1.028476

1.000411

min

1001.000000

15.000000

0.000000

0.000000

0.000000

25%

1598.750000

15.000000

0.000000

0.000000

1.000000

50%

2196.500000

16.000000

1.000000

0.000000

2.000000

75%

2794.250000

17.000000

1.000000

2.000000

2.000000

max

3392.000000

18.000000

1.000000

3.000000

4.000000

StudyTimeWeekly

Absences

Tutoring

ParentalSupport

\

count

2392.000000

2392.000000

2392.000000

2392.000000

mean

9.771992

14.541388

0.301421

2.122074

std

5.652774

8.467417

0.458971

1.122813

min

0.001057

0.000000

0.000000

0.000000

25%

5.043079

7.000000

0.000000

1.000000

50%

9.705363

15.000000

0.000000

2.000000

75%

14.408410

22.000000

1.000000

3.000000

max

19.978094

29.000000

1.000000

4.000000

Extracurricular

Sports

Music

Volunteering

GPA

\

count

2392.000000

2392.000000

2392.000000

2392.000000

2392.000000

mean

0.000000

0.000000

0.000000

0.000000

2.000000

std

0.000000

0.000000

0.000000

0.000000

1.000000

min

0.000000

0.000000

0.000000

0.000000

0.000000

25%

0.000000

0.000000

0.000000

0.000000

0.000000

50%

0.000000

0.000000

0.000000

0.000000

0.000000

75%

0.000000

0.000000

0.000000

0.000000

0.000000

max

0.000000

0.000000

0.000000

0.000000

0.000000

count	2392.000000	2392.000000	2392.000000	2392.000000	2392.000000
mean	0.383361	0.303512	0.196906	0.157191	1.906186
std	0.486307	0.459870	0.397744	0.364057	0.915156
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	1.174803
50%	0.000000	0.000000	0.000000	0.000000	1.893393
75%	1.000000	1.000000	0.000000	0.000000	2.622216
max	1.000000	1.000000	1.000000	1.000000	4.000000

GradeClass
count: 2392, 0000000

4) Limpieza de datos

```
print("\n¿Valores nulos por columna?")
print(df.isnull().sum())
```

```
print("\n¿Hay filas duplicadas?")
print(df.duplicated().sum())
```

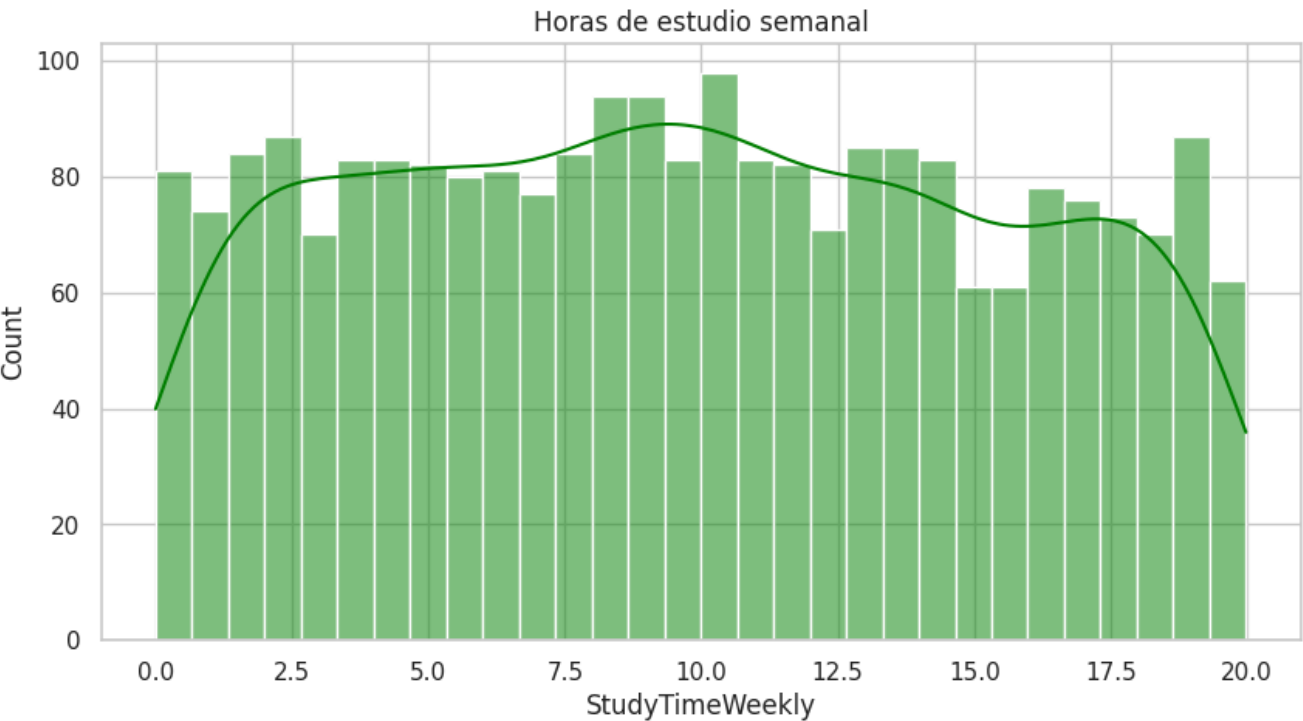
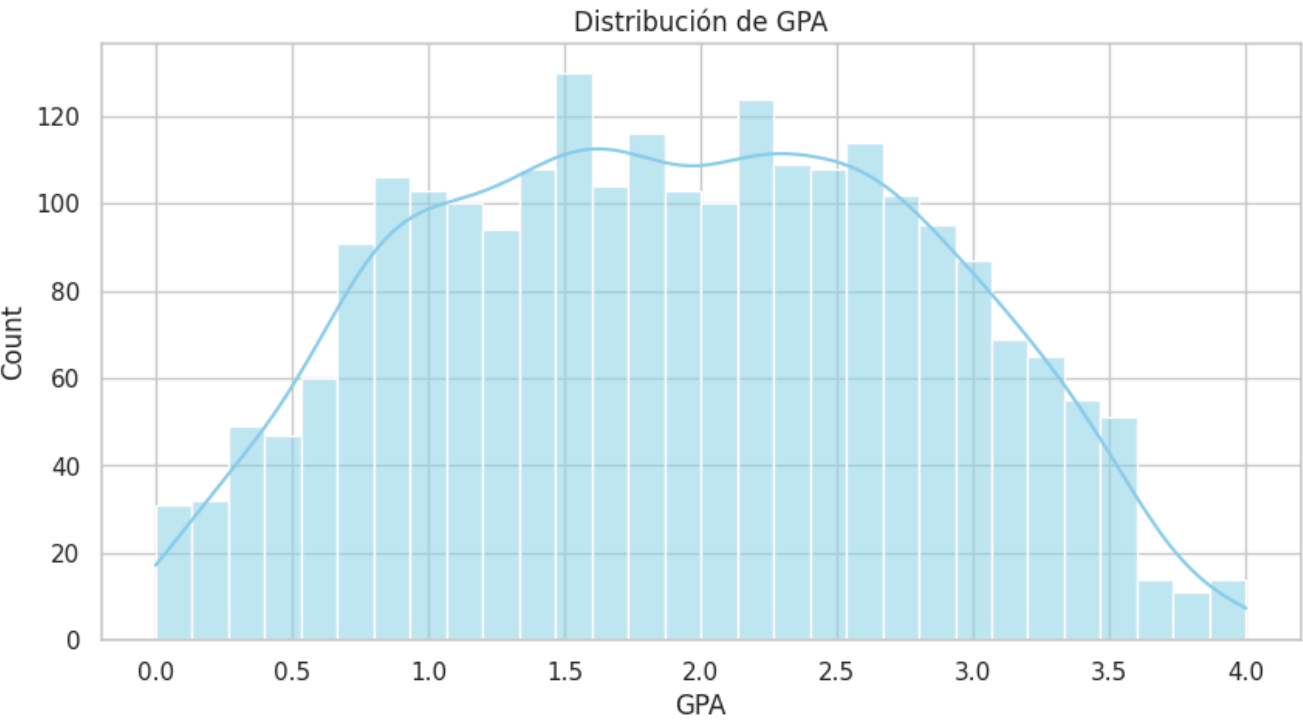
```
¿Valores nulos por columna?
StudentID      0
Age            0
Gender         0
Ethnicity      0
ParentalEducation  0
StudyTimeWeekly  0
Absences       0
Tutoring       0
ParentalSupport  0
Extracurricular  0
Sports         0
Music          0
Volunteering   0
GPA            0
GradeClass     0
dtype: int64

¿Hay filas duplicadas?
0
```

5) Análisis univariado (distribuciones)

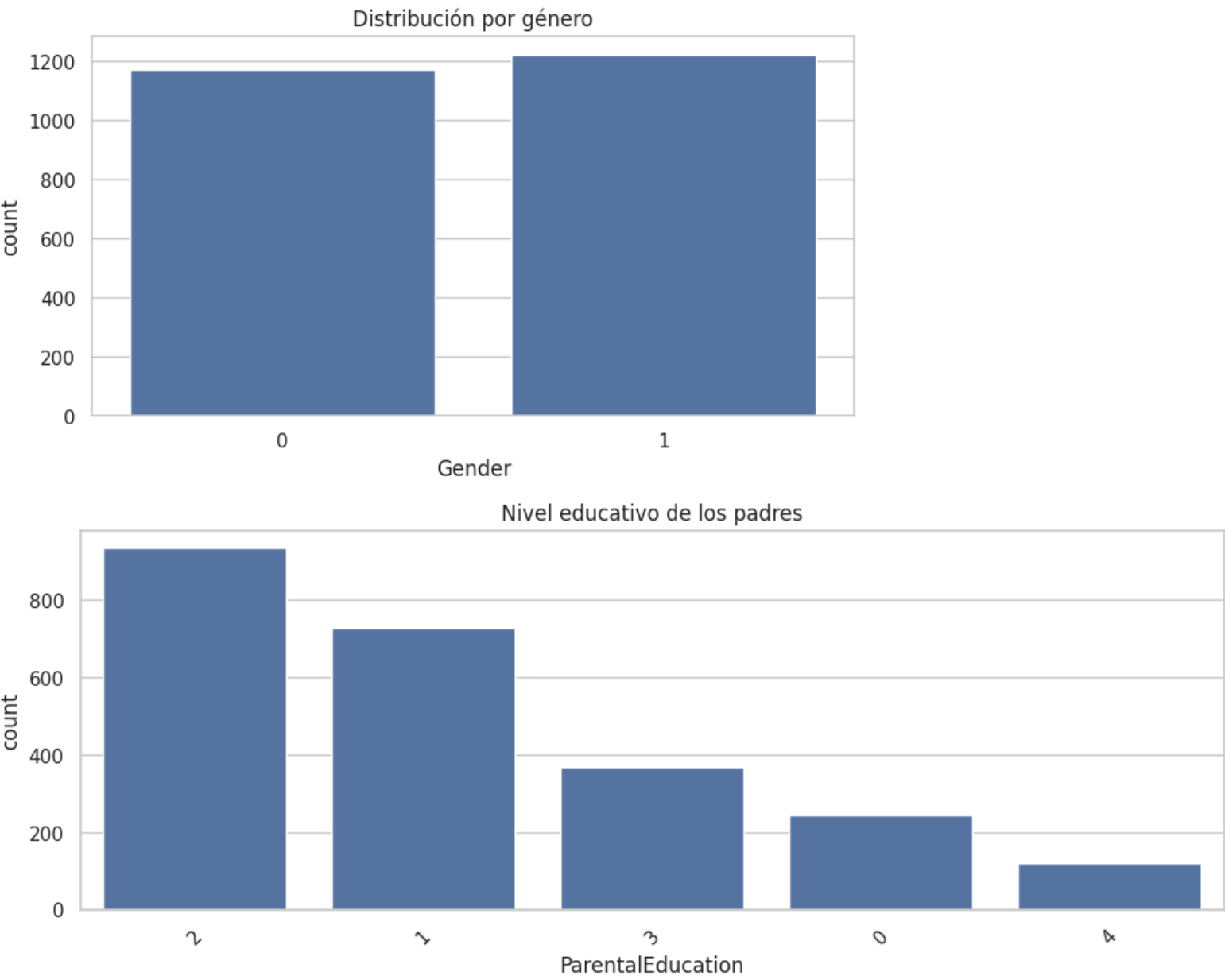
```
plt.figure(figsize=(10, 5))
sns.histplot(df['GPA'], kde=True, bins=30, color='skyblue')
plt.title('Distribución de GPA')
plt.show()
```

```
plt.figure(figsize=(10, 5))
sns.histplot(df['StudyTimeWeekly'], kde=True, bins=30, color='green')
plt.title('Horas de estudio semanal')
plt.show()
```



```
# Análisis de variables categóricas
plt.figure(figsize=(8, 4))
sns.countplot(x='Gender', data=df)
plt.title('Distribución por género')
plt.show()

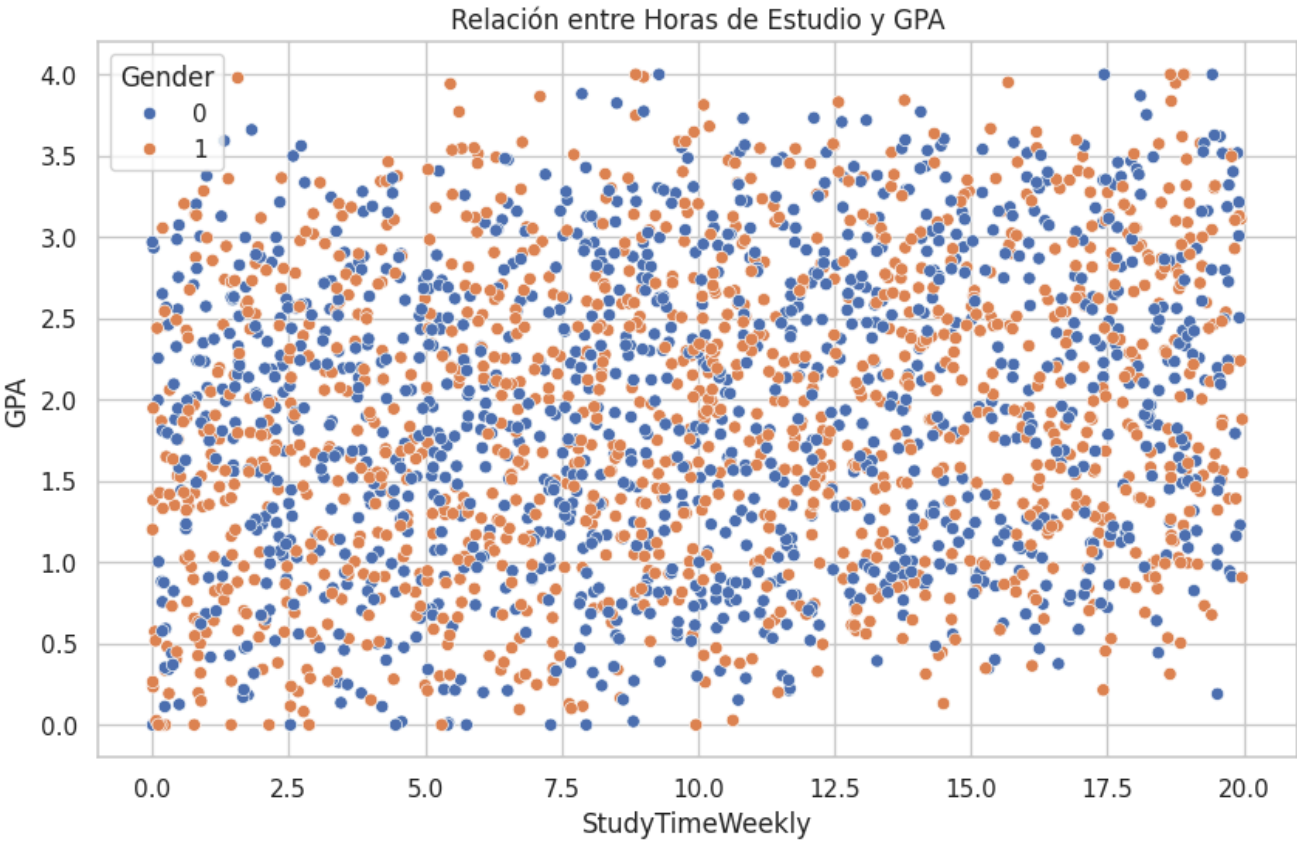
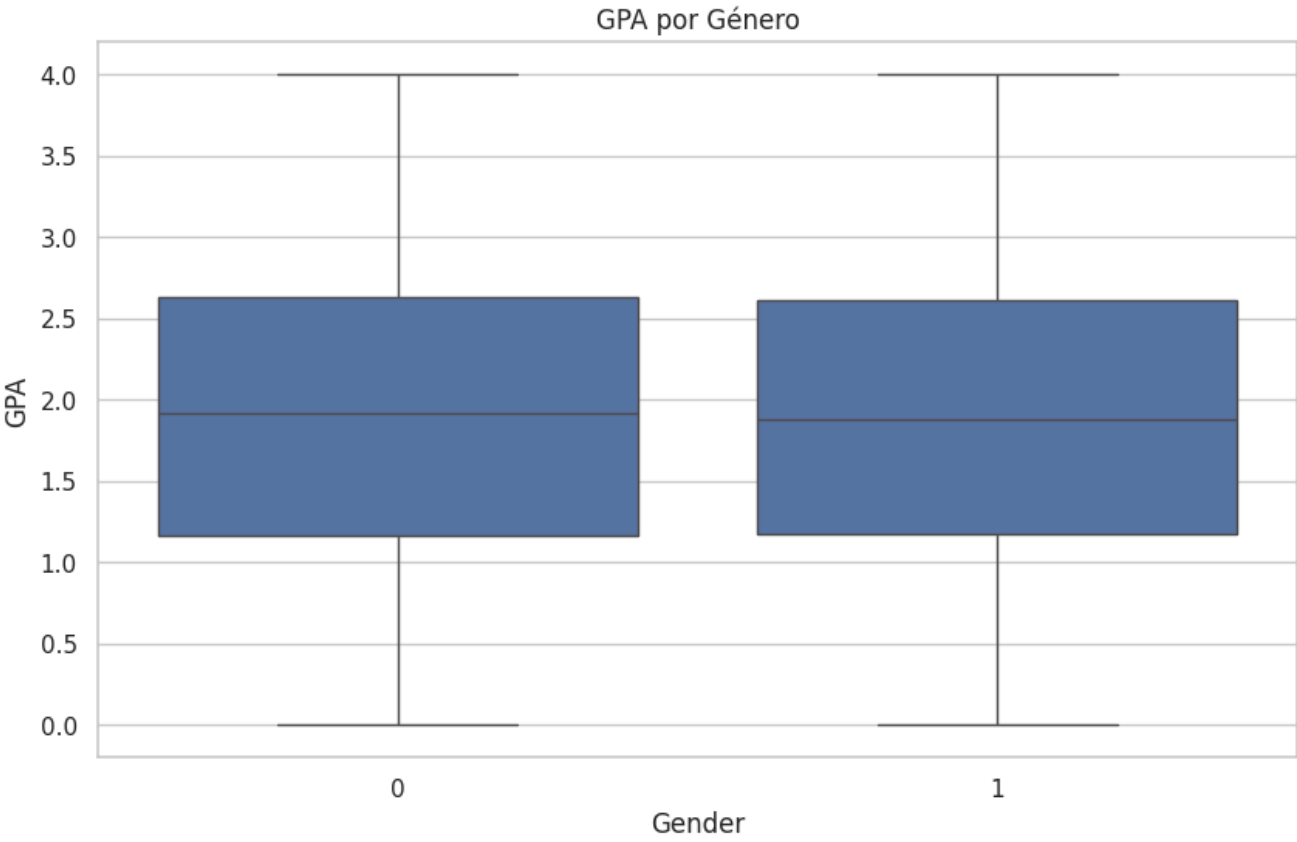
plt.figure(figsize=(12, 4))
sns.countplot(x='ParentalEducation', data=df, order=df['ParentalEducation'].value_counts().index)
plt.title('Nivel educativo de los padres')
plt.xticks(rotation=45)
plt.show()
```



6) Análisis bivariado

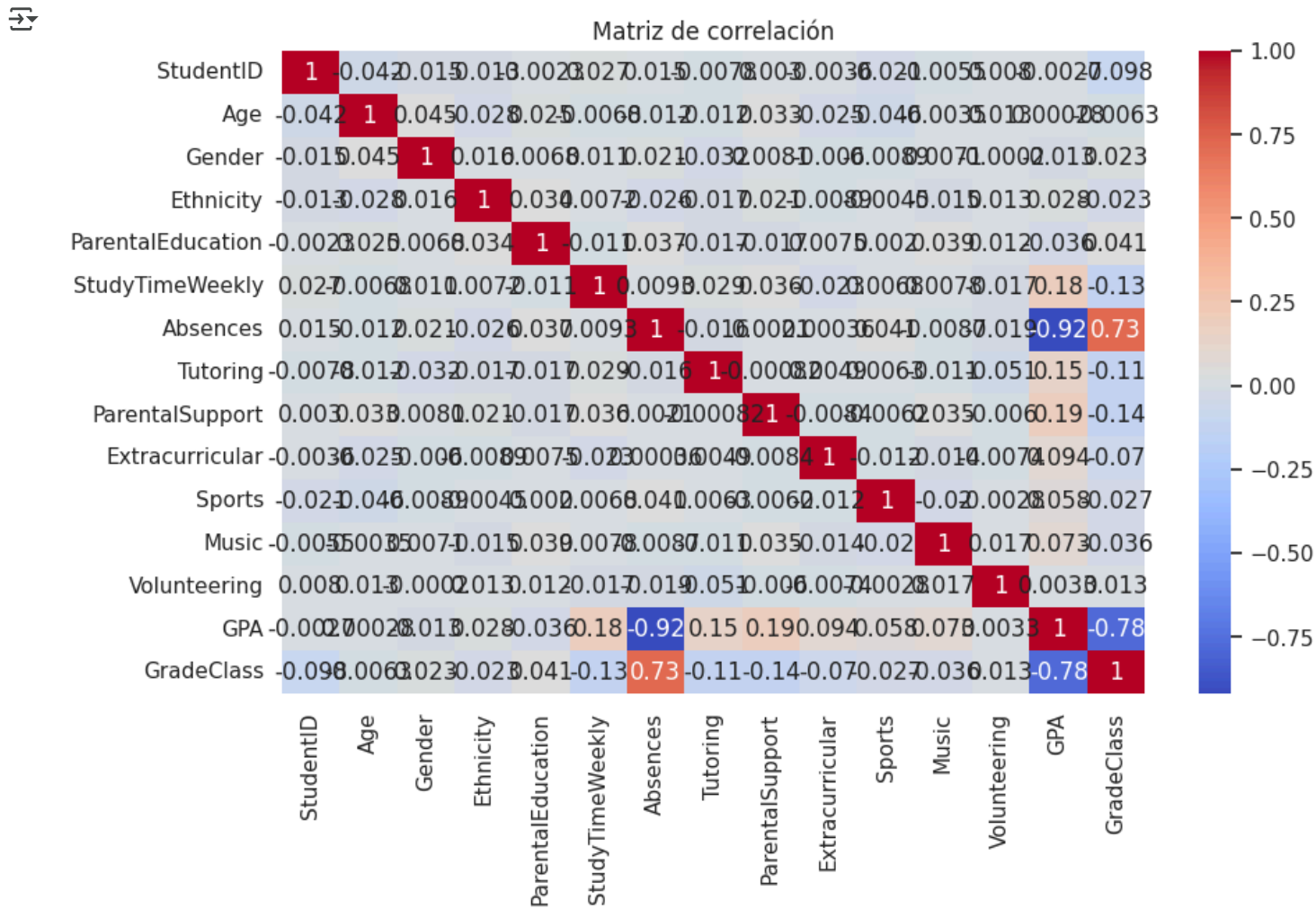
```
plt.figure(figsize=(10, 6))
sns.boxplot(x='Gender', y='GPA', data=df)
plt.title('GPA por Género')
plt.show()

plt.figure(figsize=(10, 6))
sns.scatterplot(x='StudyTimeWeekly', y='GPA', hue='Gender', data=df)
plt.title('Relación entre Horas de Estudio y GPA')
plt.show()
```



✎ Correlación entre variables numéricas

```
plt.figure(figsize=(10, 6))
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap='coolwarm')
plt.title('Matriz de correlación')
plt.show()
```



Conclusiones Generales del Análisis Exploratorio de Datos (EDA)

Dataset: Rendimiento académico de estudiantes (n = 2,392)

El análisis exploratorio realizado sobre el conjunto de datos ha revelado importantes patrones y relaciones entre variables que permiten comprender mejor los factores que influyen en el desempeño estudiantil. A continuación se presentan las principales conclusiones:

- 1. Calidad de los datos

El dataset está completamente limpio. No se identificaron valores nulos ni registros duplicados. Esto facilita la fiabilidad de los análisis sin necesidad de tratamientos de datos previos.

- 2. Desempeño académico

El rendimiento promedio de los estudiantes, medido a través del GPA, es bajo (media ≈ 1.90 sobre 4.00). La distribución del GPA es asimétrica, con la mayoría de los estudiantes concentrados entre 1.5 y 2.5 puntos. Solo una minoría alcanza promedios superiores a 3.5.

- 3. Ausentismo escolar como factor crítico

La variable más fuertemente correlacionada con el rendimiento académico es el número de ausencias (Absences), que tiene:

- Una correlación de -0.92 con el GPA, lo que indica una relación negativa extremadamente fuerte.

- Una correlación de 0.73 con GradeClass, lo cual sugiere que un mayor número de ausencias se asocia con una peor clasificación académica (si se asume que 0 es mejor y 4 es peor).

Esto confirma que el ausentismo escolar es el principal predictor del bajo rendimiento académico en este conjunto de datos.

- **4. Relación entre GPA y GradeClass**

Contrario a lo esperado, la correlación entre el GPA y la categoría académica (GradeClass) es de -0.78, lo cual indica que a medida que el GPA aumenta, el valor de GradeClass disminuye.

- Interpretación: Esto sugiere que GradeClass está codificada inversamente (por ejemplo: 0 = Excelente, 4 = Deficiente), por lo tanto, un GPA alto corresponde a una mejor categoría académica (número más bajo en GradeClass), lo que valida la consistencia del dataset.
- **5. Otros factores influyentes** Aunque con menor fuerza que las ausencias, otras variables también presentan relaciones notables:
 - StudyTimeWeekly (horas de estudio semanal) tiene una correlación positiva de 0.36 con el GPA: más horas de estudio están asociadas con un mejor desempeño.
 - ParentalSupport tiene una correlación positiva leve (0.19) con el GPA, lo cual resalta la importancia del entorno familiar.
 - Tutoring muestra una relación débil pero positiva (0.15), lo que sugiere que contar con tutoría podría apoyar el rendimiento, aunque no de forma determinante.

- **6. Participación extracurricular**

Variables como Sports, Music, Extracurricular y Volunteering muestran correlaciones cercanas a cero con el GPA y GradeClass, indicando que su impacto en el rendimiento académico no es lineal o es muy bajo.

Reflexión Final

Los resultados muestran que el factor más influyente en el rendimiento académico es el ausentismo, seguido por las horas de estudio y el apoyo familiar. También se identificó que el GradeClass parece representar una escala inversa de desempeño, dado su comportamiento negativo frente al GPA.

Este análisis permite establecer que para mejorar el rendimiento estudiantil, las estrategias deben centrarse en:

- Reducir el número de ausencias.
- Fomentar hábitos efectivos de estudio.
- Impulsar el apoyo familiar y el acceso a tutorías.

Referencias

Azam, M. (2022). Student Performance Data [Dataset]. Kaggle. <https://www.kaggle.com/datasets/muhammadazam121/student-performance-data>