Universidad Cooperativa de Colombia

Proyecto Final

Andrés Chacua

Juan Patiño

Gabriel Narvaez

San Juan De Pasto

2 de mayo de 2025

[Introducción 1](#_Toc963392681)

[Metodología 1](#_Toc178151584)

[Desarrollo del proyecto 1](#_Toc489901163)

[Gráficos 2](#_Toc1157657192)

[Histograma con curva de densidad 2](#_Toc1520512060)

[Diagrama de caja 2](#_Toc2017994540)

[Gráfico Q-Q 3](#_Toc1823572201)

[Género (Choose.your.gender.) 3](#_Toc769327507)

[Tabla de frecuencias 4](#_Toc882246744)

[Porcentajes de cada género respecto al total 4](#_Toc795834468)

[Gráficos 4](#_Toc420009411)

[Promedio Académico (What.is.your.CGPA.) 5](#_Toc1149422619)

[Tabla de frecuencias 5](#_Toc1601497714)

[Porcentajes para el rango de notas respecto al total 5](#_Toc663105135)

[Gráficos 6](#_Toc2128118172)

[Presenta depresión (Do.you.have.Depression.) 6](#_Toc2123718276)

[Tabla de frecuencias 6](#_Toc1275444599)

[Porcentaje respecto a si presenta o no presenta depresión 6](#_Toc1719348483)

[Gráficos 6](#_Toc842155161)

[Presenta Ansiedad (Do.you.have.Anxiety.) 7](#_Toc271364443)

[Porcentaje respecto a si presenta o no presenta ansiedad 8](#_Toc651451132)

[Gráficos 8](#_Toc633376866)

[Presenta ataques de Pánico (Do.you.have.Panic.atack.) 9](#_Toc2111867201)

# Introducción

Este proyecto tiene como objetivo aplicar los diferentes temas estudiados durante el curso de probabilidad y estadística, basándonos en un conjunto de datos relacionados con la salud mental de estudiantes de diferentes carreras universitarias.

El estudio de estos datos abarca diferentes variables como la edad, género, su promedio académico (CGPA), la presencia de diferentes condiciones como la ansiedad, depresión o ataques de pánico.

Desde diferentes herramientas estadísticas se analizarán los diferentes datos, basándonos en el comportamiento de cada uno, además identificaremos posibles distribuciones de probabilidad, sacar conclusiones sobre la población y examinar relaciones entre las variables más relevantes.

De esta manera, buscamos entender mejor los factores que podrían impactar en la salud mental dentro del contexto universitario.

## Metodología

Para poder llevar acabo el desarrollo de este proyecto utilizamos un conjunto de datos elegidos desde la página web de Kaggle, estos datos se basaron en encuestas hacia estudiantes universitarios, los cuales contienen variables como la edad, género, salud mental o promedio académico.

El análisis de estos datos se desarrolló a través del software Rstudio y diferentes librerías como readr, dplyr, para el manejo, análisis y la visualización de los datos a través de gráficos.

Las variables para para el análisis estadístico son las siguientes:

* Edad: variable cuantitativa continúa
* Género: variable cualitativa
* Promedio Académico (CGPA): variable cuantitativa discreta
* Presencia de ansiedad: variable cualitativa
* Presencia de depresión: variable cualitativa
* Presencia de ataques de pánico: variable cualitativa

Posterior a ello, se realizaron las siguientes actividades:

* Análisis descriptivo de las variables mediante medidas de tendencia central y dispersión.
* Análisis de distribución para determinar si las variables cuantitativas siguen una distribución conocida.
* Inferencia estadística mediante estimaciones de parámetros y pruebas de hipótesis.
* Análisis de correlación y regresión lineal simple.

## Desarrollo del proyecto

1. **Estadística descriptiva**

En este punto analizamos cada variable del estudio de manera individual, utilizando medidas de tendencia central, dispersión y gráficos para poder ver como se comportan estos datos.

**Edad (Age)**

**Media, mediana, mínimo, máximo:**

* Mínimo: 18 años
* 1er cuartil (Q1): 18 años
* Mediana (Q2): 19 años
* Media: 20.53
* 3cer cuartil (Q3): 23 años
* Máximo: 24 años

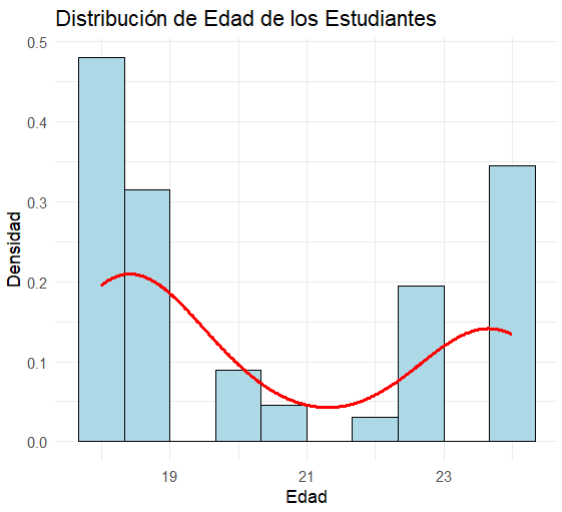
**Desviación estándar**: 2.49628

**Varianza**: 6.231414

**Rango**: 18, 24

## Gráficos

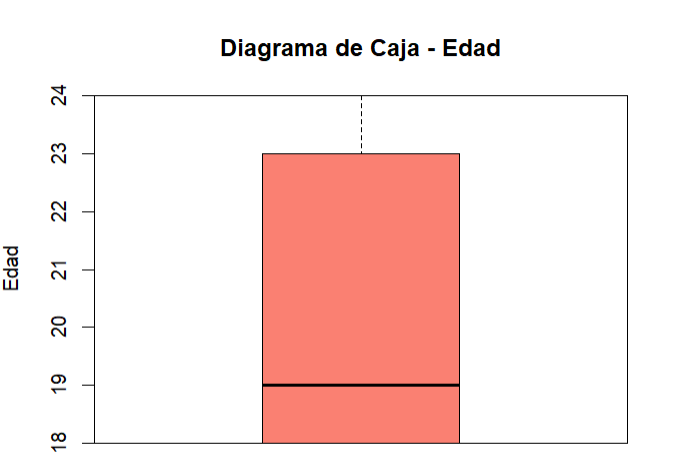
Histograma con curva de densidad

****

**Análisis**

* La mayoría de los estudiantes tiene unos 18 años, pero también hay un grupo que rondan los 24. Parece que hay dos picos, uno cerca de los 18 y otro cerca de los 24. Esto no señala que hay una ola juventud, pero también estudiantes más grandes que todavía están en la universidad en la que se realizó la encuesta de nuestra base de datos.

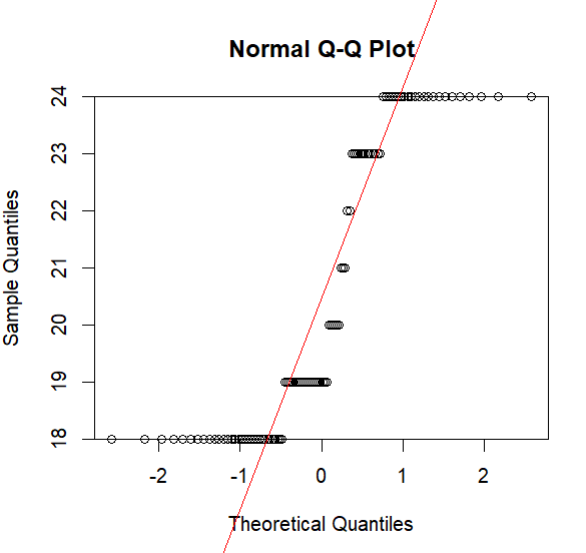
Diagrama de caja



Analisis

* La mitad de la banda tiene 18 o 19 años, y la mayoría está entre los 21. Solo unos pocos pasan los 22, así que la mayoría son jóvenes. No hay datos raros o extremos, todo parece normal.

Gráfico de dispersión

****

**Analisis**

* La edad no sigue exactamente una distribución normal, pero está bastante cerca. Algunos puntos no encajan perfecto, pero básicamente, la edad de los estudiantes está bastante equilibrada y no hay cosas muy fuera de lugar.

## Género (Choose.your.gender.)

Tabla de frecuencias

|  |  |  |
| --- | --- | --- |
|  | Variable | Freq |
| 1. | Femenino | 75 |
| 2. | Masculino | 26 |

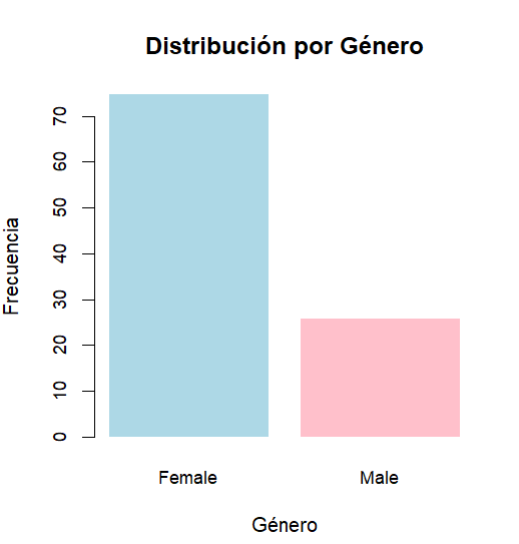
### Porcentajes de cada género respecto al total

Femenino: 74.25743%

Masculino: 25.74257%

### Gráficos

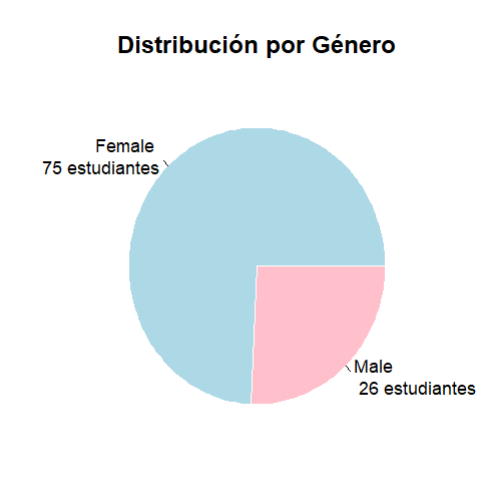
1. Diagrama de barras



Analisis

* La mayoría son más mujeres, como tres veces más que los hombres. O sea, en la cantidad de alumnos que respondieron esta encuesta hay muchas más mujeres que hombres, esto influye en cómo se ve todo en general.

1. Gráfico circular



## Promedio Académico (What.is.your.CGPA.)

**Rangos más comunes**:

* 3.00 – 3.49: 43 estudiantes
* 3.50 – 4.00: 47 estudiantes

Tabla de frecuencias

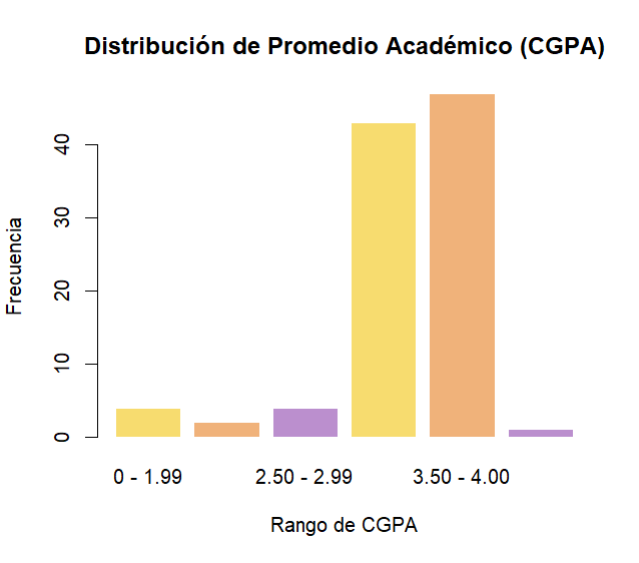


### Porcentajes para el rango de notas respecto al total

* + 0 – 1.99: 3.96039%
  + 2.00 – 2.49: 1.98019%
  + 2.50 – 2.99: 3.96039%
  + 3.00 – 3.49: 42.57425%
  + 3.50 – 4.00: 46.53465%
  + 3.50 – 4.00: 0.99009%

### Gráficos

1. Diagrama de barras



Analisis

* La mayoría de los estudiantes tienen un promedio entre 3.50 y 4.00, lo que indica un alto nivel académico en general. La frecuencia en ese rango es mucho mayor comparada con otros rangos de CGPA, así que podemos decir que en ese grupo hay muchos estudiantes que mantienen buenas calificaciones.

## Presenta depresión (Do.you.have.Depression.)

Tabla de frecuencias

|  |  |  |
| --- | --- | --- |
|  | Variable | Freq |
| 1 | No | 66 |
| 2 | Sí | 35 |

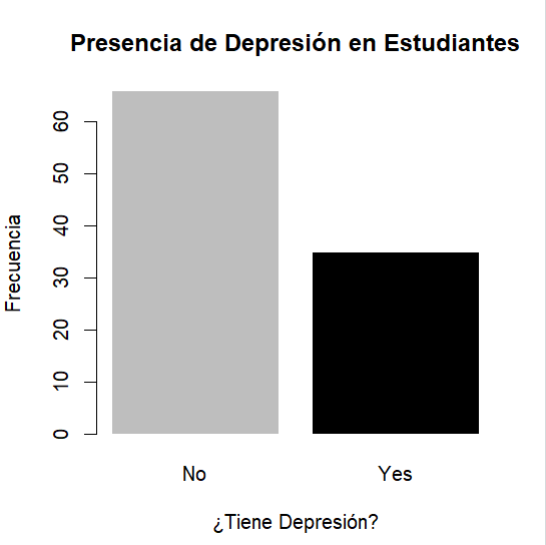
### Porcentaje respecto a si presenta o no presenta depresión

No: 65.34653%

Sí: 34.65347

### Gráficos

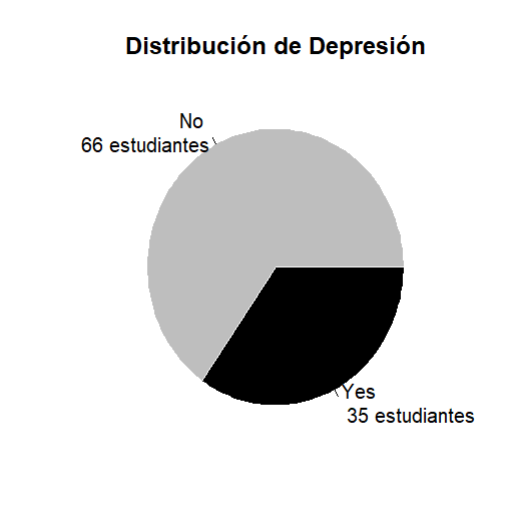
1. Diagrama de barras



Análisis

La mayoría de los estudiantes (más de 60) no presenta depresión, mientras que una cantidad menor (unos 35) sí la padece. Esto indica que, aunque la depresión no es algo que afecte a toda la muestra, sí hay un porcentaje importante de estudiantes que experimenta esta condición. Es relevante porque muestra que la depresión es un tema presente en la población estudiada, pero no predominante.

1. Diagrama circular



## Presenta Ansiedad (Do.you.have.Anxiety.)

**Tabla de frecuencias**

|  |  |  |
| --- | --- | --- |
|  | Variable | Freq |
| 1 | No | 67 |
| 2 | Sí | 34 |

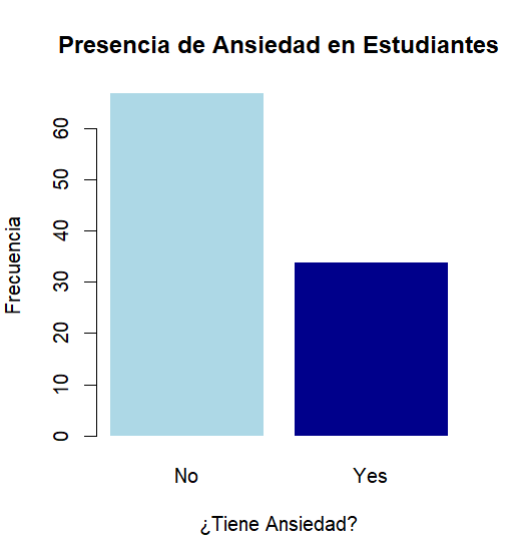
### Porcentaje respecto a si presenta o no presenta ansiedad

No: 66.33663 33.66337

Sí: 33.66337

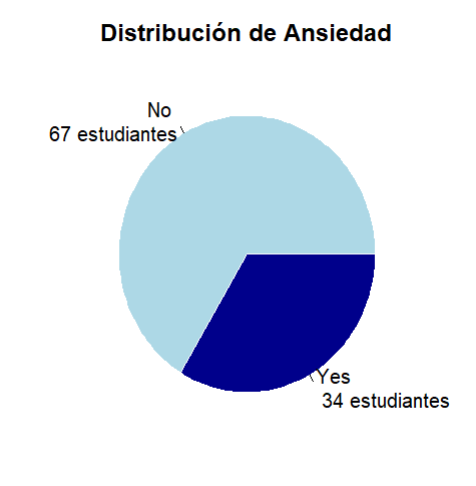
### Gráficos

1. Diagrama de barras



* La gráfica muestra que más de la mitad de los estudiantes no experimenta ansiedad. La cantidad de estudiantes que dice que siente ansiedad es menor, lo que indica que la ansiedad es un tema bastante común en esta muestra.

1. Diagrama circular

****

## Presenta ataques de Pánico (Do.you.have.Panic.atack.)

**Tabla de frecuencias**

|  |  |  |
| --- | --- | --- |
|  | Variable | Freq |
| 1 | No | 68 |
| 2 | Sí | 33 |

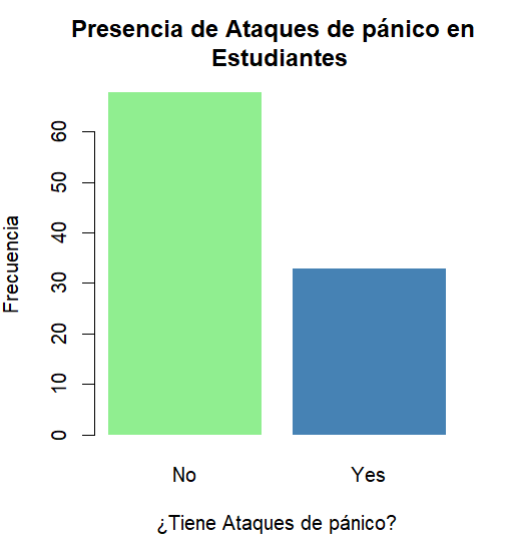
**Porcentaje respecto a si presenta o no presenta ataques de pánico**

No: 67.32673

Sí: 32.67327

### Gráficos

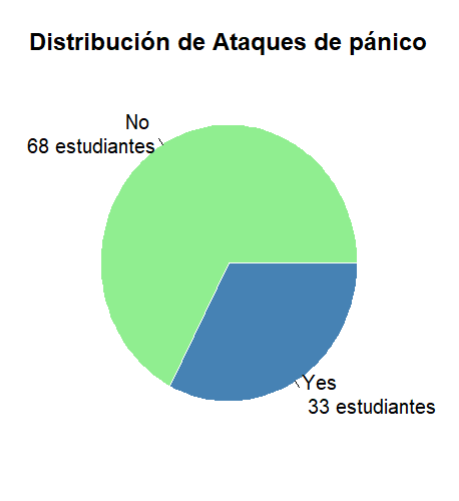
1. Diagrama de barras



**Análisis**

La mayoría de los estudiantes no sufre ataques de pánico. Sin embargo, hay un porcentaje importante (unos 33 estudiantes) que sí los experimenta, por lo que no es un tema menor. La gráfica circular confirma que hay un grupo significativo que lidia con ataques de pánico, lo cual puede ser importante para tomar medidas o investigaciones sobre salud mental.

1. Diagrama circular

****

**Análisis general**

En conjunto, los datos reflejan una población estudiantil mayoritariamente joven, con mayor presencia femenina, con buen rendimiento académico y sin niveles elevados de ansiedad, depresión ni ataques de pánico, sin embargo, hay un grupo significativo que sí enfrenta estos problemas de salud mental, especialmente ansiedad y depresión. Esto indica que, además de centrarse en el rendimiento académico, también es importante prestar atención al bienestar emocional y psicológico de los estudiantes.

1. **Distribuciones de probabilidad**

**Shapiro wilk**

Para analizar las distribuciones se tomó la variable Edad (Age), ya que es una variable continúa numérica, esto, con el objetivo de determinar Age sigue una distribución normal. Para llevar a cabo esto aplicamos la prueba de Shapiro-Wilk, obteniendo los siguientes resultados:

Data: Age

W = 0.78617, p-value = 9.519e-11

Los datos nos muestran que p-valor es menor a 0.05, esto rechaza la hipótesis nula de que los datos se derivan de una distribución normal. Esto lo podemos confirmar visualmente a través del gráfico de histograma y el gráfico Q-Q mostrados anteriormente, en los cuales, estos muestran una distribución irregular y alejada a la simetría y alejada de la simetría característica de una distribución que suele ser normal.

**Chi-Cuadrado**

Se aplicó una prueba de Chi-cuadrado de independencia para evaluar si existe una relación entre el género del estudiante y la presencia de síntomas de depresión.

El resultado de la prueba fue:

X-squared = 1.4409, df = 1, p-value = 0.23

Interpretación:

Dado que el p-valor es mayor a 0.05, no se rechaza la hipótesis nula.

Por lo tanto, no hay evidencia estadísticamente significativa para afirmar que exista una asociación entre el género del estudiante y la presencia de depresión en esta muestra.

1. Inferencia estadística

**Prueba de Hipótesis: Edad vs Depresión**

Se aplicó una **prueba t para muestras independientes** para evaluar si existe una diferencia significativa en la edad promedio entre estudiantes que tienen y no tienen síntomas de **depresión**.

**Resultado:**

* **Media edad (sin depresión):** 20.66 años
* **Media edad (con depresión):** 20.29 años
* **t = 0.7306**, **df = 73.77**, **p-value = 0.4673**
* **IC 95% de la diferencia de medias:** (-0.65, 1.40)

**Interpretación:**  
Dado que el **p-valor es mayor a 0.05**, **no se rechaza la hipótesis nula**.  
Esto indica que **no hay evidencia estadísticamente significativa** para afirmar que existe una diferencia en la edad promedio entre estudiantes con y sin depresión.

**Prueba de Hipótesis: Edad vs Ansiedad**

También se aplicó una prueba t para determinar si existe diferencia significativa en la edad promedio entre estudiantes que presentan **ansiedad** y los que no.

**Resultado:**

* **Media edad (sin ansiedad):** 20.70 años
* **Media edad (con ansiedad):** 20.21 años
* **t = 0.9295**, **df = 66.41**, **p-value = 0.356**
* **IC 95% de la diferencia de medias:** (-0.56, 1.54)

**Interpretación:**  
Nuevamente, dado que el **p-valor es mayor a 0.05**, **no se rechaza la hipótesis nula**.  
Por lo tanto, **no se encontró una diferencia significativa** en la edad promedio entre estudiantes con y sin ansiedad.

Prueba de Hipótesis: CGPA en mujeres (con vs sin depresión)

Se aplicó una **prueba t para muestras independientes**, considerando únicamente a las **estudiantes mujeres**, con el fin de evaluar si existe una diferencia significativa en el **promedio académico (CGPA)** entre quienes presentan síntomas de **depresión** y quienes no.

**Resultado:**

* **Media CGPA (sin depresión):** 3.4327
* **Media CGPA (con depresión):** 3.3898
* **t = 0.4283**, **df = 71.92**, **p-value = 0.6697**
* **IC 95% de la diferencia de medias:** (-0.1568, 0.2426)

**Interpretación:**  
Dado que el **p-valor es mayor a 0.05**, **no se rechaza la hipótesis nula**.  
Esto indica que **no hay evidencia estadísticamente significativa** para afirmar que existe una diferencia en el promedio académico entre mujeres con y sin depresión.  
La diferencia de medias es baja (alrededor de 0.04 puntos), y el intervalo de confianza incluye el 0, lo que refuerza esta conclusión.

Prueba de Hipótesis: CGPA en hombres (con vs sin depresión)

También se realizó la misma prueba t con los **estudiantes hombres**.

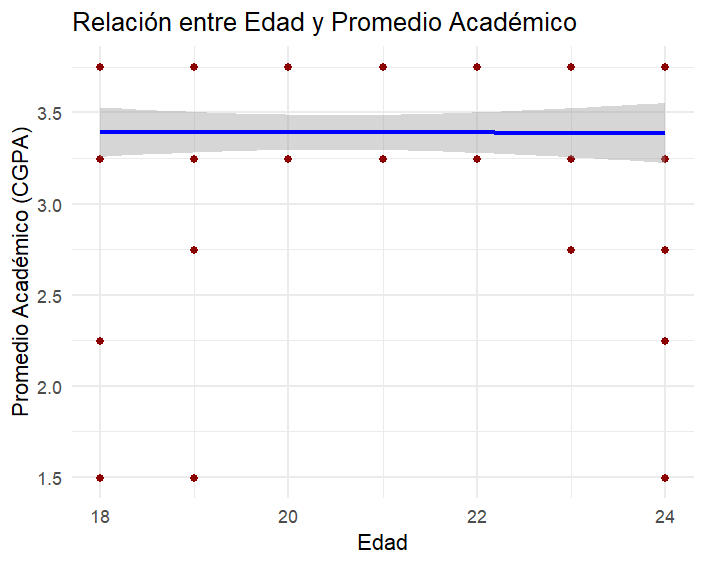
**Resultado:**

* **Media CGPA (sin depresión):** 3.2220
* **Media CGPA (con depresión):** 3.3292
* **t = -0.6070**, **df = 23.95**, **p-value = 0.5496**
* **IC 95% de la diferencia de medias:** (-0.4716, 0.2573)

**Interpretación:**  
El **p-valor también es mayor a 0.05**, por lo tanto, **no se rechaza la hipótesis nula**.  
Esto indica que **no existe una diferencia significativa** en el promedio académico entre hombres con y sin depresión.  
Aunque la media fue ligeramente mayor en los hombres con depresión, la diferencia es pequeña y estadísticamente no significativa.

1. Regresión o correlación

Relación entre edad y promedio académico



El gráfico muestra que no existe una relación significativa entre la edad y el promedio académico (CGPA) de los estudiantes. La línea de regresión es prácticamente horizontal, lo que indica que, al aumentar la edad, el promedio académico se mantiene casi constante. Además, los puntos están agrupados en niveles discretos, reflejando que los valores de CGPA provienen de rangos predefinidos. El intervalo de confianza alrededor de la línea es estrecho, lo que sugiere precisión en la estimación, pero no implica una relación significativa. En resumen, la edad no parece influir de manera relevante en el rendimiento académico según los datos disponibles.

¿Como afecta la depresión en el promedio académico según la edad del estudiante?

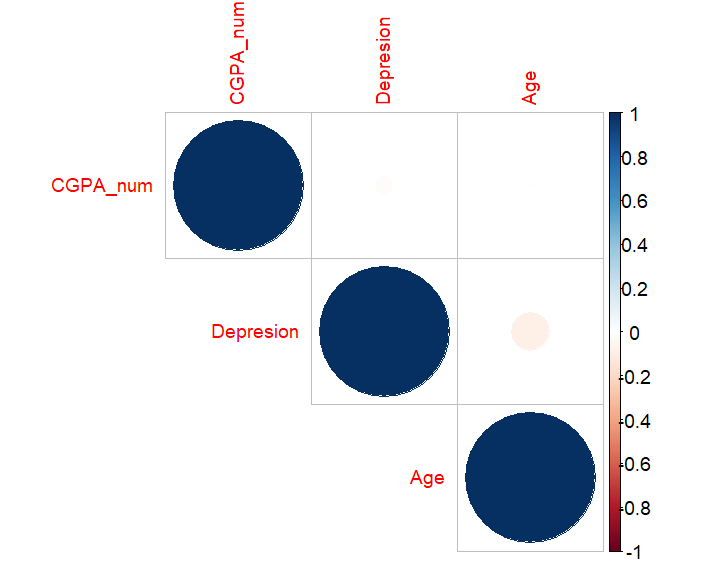
Matriz de correlación de Pearson

|  |  |  |  |
| --- | --- | --- | --- |
|  | CGPA\_num | Depresion | Age |
| CGPA\_num | 1.000 | -0.0187 | -0.0038 |
| Depresion | -0.0187 | 1.000 | -0.0871 |
| Age | -0.0038 | -0.0871 | 1.000 |

Interpretación

* CGPA y Depresión: -0.0187 → Correlación casi nula, no hay relación.
* CGPA y Edad: -0.0038 → Cero relación.
* Edad y Depresión: -0.0871 → Relación débil negativa, pero prácticamente nula también.

En resumen: no hay correlaciones significativas entre estas variables.



El gráfico de correlación muestra que no existe una relación fuerte entre las variables CGPA\_num (promedio académico), Depresión y Age (edad), ya que los círculos fuera de la diagonal principal son pequeños y de colores pálidos, lo que indica valores de correlación cercanos a cero. Esto sugiere que el promedio académico no se asocia significativamente ni con la edad ni con la presencia de depresión, y que la edad tampoco está relacionada de manera importante con la depresión. En general, las variables parecen ser independientes entre sí dentro de este conjunto de datos.

1. Conclusiones

1. Usar RStudio nos permitió ir más allá de simplemente mirar tablas de datos: aprendimos a “conversar” con los datos. Por ejemplo, al aplicar una prueba t para comparar el promedio académico entre hombres y mujeres, no solo se aprendió a ejecutar la fórmula, sino también a interpretar qué significa un valor p y cómo usarlo para tomar decisiones reales. Esto nos mostró que RStudio no es solo para estadísticos, sino también para ingenieros que quieren entender mejor su entorno a través de la información.

2. La generación de gráficos en RStudio (como los diagramas de dispersión, histogramas y matrices de correlación) fue clave para visualizar relaciones entre variables. Un detalle puntual muy valioso fue darnos cuenta de cómo se puede ver de forma inmediata si una variable está relacionada con otra, como al graficar los niveles de depresión frente al promedio académico. Esto no solo facilita la interpretación, sino que también ayuda a comunicar hallazgos a personas que no necesariamente dominan los datos. En otras palabras, aprendimos que una buena gráfica puede decir más que una tabla llena de números.

3. La clase de Probabilidad y Estadística, vista desde la perspectiva de un ingeniero de software, va mucho más allá de resolver ejercicios de distribuciones o calcular medias. Nos enseñó a pensar con lógica y a tomar decisiones con respaldo numérico. Por ejemplo, en el desarrollo de aplicaciones, esta base sirve para analizar el comportamiento de usuarios, medir el impacto de una nueva función o predecir fallas. Un punto interesante fue ver cómo conceptos como varianza y correlación se usan realmente en áreas como machine learning o análisis de rendimiento. En definitiva, la estadística no solo es útil, es parte esencial del toolkit de un buen ingeniero.