Taller de Análisis Exploratorio de Datos (EDA)

con R Studio

Predicción del Rendimiento Académico Estudiantil

Colegio Universitario de Cartago
Curso: Introducción a la Estadística / Probabilidad y Estadística

Profesor: David Martínez Salazar
III Cuatrimestre 2025

1 Introducción al Taller

1.1 Objetivo General

En este taller práctico, se trabajará con un conjunto de datos real que contiene registros académicos de **2000 estudiantes**. El estudiante aprenderá a realizar un **Análisis Exploratorio de Datos (EDA)** utilizando R Studio, desde la importación de datos hasta la construcción de un modelo de regresión lineal simple.

1.2 ¿Qué es el EDA?

✦ Concepto Teórico

El **Análisis Exploratorio de Datos (EDA)** es el proceso de examinar y resumir las características principales de un conjunto de datos. Sus objetivos son:

- Comprender la estructura y contenido de los datos
- Detectar errores, valores faltantes o atípicos
- Calcular estadísticas descriptivas (media, mediana, moda, etc.)
- Visualizar patrones mediante gráficos
- Verificar supuestos estadísticos (como la normalidad)
- Identificar relaciones entre variables

1.3 Objetivos Específicos

Al finalizar este taller, el estudiante será capaz de:

1. Importar y explorar datos en R Studio
2. Calcular medidas de tendencia central y dispersión
3. Crear gráficos estadísticos (pastel, barras, histogramas, boxplots)
4. Evaluar la normalidad de una variable
5. Analizar correlación entre dos variables
6. Construir un modelo de regresión lineal simple

1.4 Modalidad de Trabajo

- **Modalidad:** Individual o en parejas
- **Herramienta:** R Studio
- **Duración estimada:** 2-3 horas
- **Entregable:** Script de R (.R) con todos los análisis

2 Descripción del Conjunto de Datos

2.1 Contexto del Dataset

El dataset **"Student Performance Prediction"** fue obtenido de la plataforma Kaggle y contiene información académica de **2000 estudiantes**. Este conjunto de datos permite analizar y predecir las **calificaciones del examen final** basándose en diversos indicadores académicos.

✦ Concepto Teórico

¿Por qué es importante este análisis?
En el ámbito educativo, identificar los factores que influyen en el rendimiento académico permite:

- Diseñar intervenciones tempranas para estudiantes en riesgo
- Optimizar estrategias de enseñanza-aprendizaje
- Comprender qué hábitos conducen al éxito estudiantil
- Desarrollar sistemas de alerta temprana basados en datos

2.2 Variables del Dataset

El dataset contiene las siguientes variables:

Variable Original	Tipo	Descripción
`Student_ID`	Identificador	Código único del estudiante
`Attendance_Rate`	Numérica (0-100)	Porcentaje de asistencia a clases
`Internal_Test_1`	Numérica (0-100)	Calificación del primer examen interno
`Internal_Test_2`	Numérica (0-100)	Calificación del segundo examen interno
`Assignment_Avg`	Numérica (0-100)	Promedio de calificaciones en tareas
`Study_Hours_Per_Day`	Numérica	Horas diarias dedicadas al estudio
`Final_Exam_Mark`	Numérica (0-100)	Calificación del examen final

Table 1: Descripción de las variables del dataset

2.3 Flujo del Análisis EDA

En este taller se seguirá el siguiente orden:

1. **Importación de datos:** Cargar el dataset en R
2. **Exploración inicial:** Ver estructura y primeros registros
3. **Limpieza de datos:** Renombrar columnas, verificar valores faltantes

4. **Estadística descriptiva:** Calcular media, mediana, moda, etc.
5. **Visualización:** Gráficos de pastel, barras y boxplots
6. **Análisis de normalidad:** Histograma, Q-Q plot, pruebas estadísticas
7. **Correlación:** Relación entre asistencia y nota final
8. **Regresión lineal:** Modelo predictivo simple

3 Configuración del Entorno de Trabajo

✔ Instrucciones

Antes de comenzar, se deben instalar y cargar las librerías necesarias. Es importante seguir cada paso cuidadosamente.

3.1 Paso 1: Instalación de Paquetes

★ Nota Importante

Solo es necesario instalar los paquetes UNA VEZ. Si ya se encuentran instalados, se puede omitir este paso.

```r
# Instalar paquetes necesarios
install.packages("ggplot2")    # Para graficos
install.packages("dplyr")      # Para manipular datos
install.packages("corrplot")   # Para mapa de calor
install.packages("reshape2")   # Para reestructurar datos
```

Listing 1: Instalación de paquetes (ejecutar solo una vez)

3.2 Paso 2: Cargar las Librerías

```r
# Cargar librerias
library(ggplot2)
library(dplyr)
library(corrplot)
library(reshape2)

# Mostrar numeros sin notacion cientifica
options(scipen = 999)
```

Listing 2: Cargar librerías al inicio de cada sesión

3.3 Paso 3: Importar el Dataset

```r
# Importar el dataset
# NOTA: Se debe ajustar la ruta segun la ubicacion del archivo
datos <- read.csv("student_performance.csv")

# Verificar que se cargo correctamente
cat("El dataset tiene", nrow(datos), "filas y", ncol(datos),
    "columnas\n")
```

Listing 3: Importar el archivo CSV

✎ Actividad

Se debe ejecutar el código anterior y anotar cuántas filas y columnas tiene el dataset.

✍ Espacio para Respuesta

Número de filas: _____

Número de columnas: _____

4 Exploración Inicial de los Datos

4.1 Ver las Primeras y Últimas Filas

```r
# Ver las primeras 6 filas
head(datos)

# Ver las ultimas 6 filas
tail(datos)
```

Listing 4: Explorar las primeras filas del dataset

4.2 Estructura del Dataset

```r
# Ver estructura: tipos de variables
str(datos)

# Ver nombres de las columnas
names(datos)
```

Listing 5: Ver la estructura de los datos

✎ Actividad

Se debe ejecutar el código `str(datos)` y responder:

1. ¿Cuántas variables numéricas existen?
2. ¿Existe alguna variable de tipo texto (character)?

✍ Espacio para Respuesta

1. Variables numéricas: _____

2. Variables de texto: _____

4.3 Resumen Estadístico Rápido

```r
# Resumen de todas las variables
summary(datos)
```

Listing 6: Resumen estadístico general

✎ Actividad

Con el resultado de `summary(datos)`, se deben anotar los valores para la variable **Final_Exam_Mark**:

✍ **Espacio para Respuesta**

Mínimo: _____ **Máximo:** _____

Mediana: _____ **Media:** _____

5 Limpieza y Preparación de Datos

5.1 Renombrar las Columnas

Para facilitar el trabajo, se renombrarán las columnas a español.

```r
# Ver nombres actuales
names(datos)

# Renombrar las columnas
colnames(datos) <- c(
  "ID",                    # Student_ID
  "Asistencia",            # Attendance_Rate
  "Examen_Interno_1",      # Internal_Test_1
  "Examen_Interno_2",      # Internal_Test_2
  "Promedio_Tareas",       # Assignment_Avg
  "Horas_Estudio",         # Study_Hours_Per_Day
  "Nota_Final"             # Final_Exam_Mark
)

# Verificar los nuevos nombres
names(datos)
```

Listing 7: Renombrar columnas a español

5.2 Verificar Valores Faltantes

```r
# Contar valores NA en cada columna
colSums(is.na(datos))

# Total de valores faltantes
cat("Total de valores faltantes:", sum(is.na(datos)), "\n")
```

Listing 8: Buscar valores faltantes (NA)

✎ Actividad

¿Existen valores faltantes en el dataset? ¿En cuáles variables?

✍ Espacio para Respuesta

¿Existen valores faltantes? (Sí / No): _____

¿En cuáles variables? _____

6 Estadística Descriptiva: Nota del Examen Final

El análisis se enfocará en la variable **Nota_Final** (calificación del examen final).

✦ Concepto Teórico

Medidas de Tendencia Central:

- **Media ($\bar{x}$):** Promedio de todos los valores
- **Mediana:** Valor que divide los datos a la mitad
- **Moda:** Valor que más se repite

Medidas de Dispersión:

- **Varianza (s^2):** Indica qué tan dispersos están los datos
- **Desviación Estándar (s):** Raíz cuadrada de la varianza
- **Cuartiles:** Dividen los datos en 4 partes iguales

6.1 Medidas de Tendencia Central

```r
# MEDIA (promedio)
media <- mean(datos$Nota_Final)
cat("Media:", round(media, 2), "\n")

# MEDIANA (valor central)
mediana <- median(datos$Nota_Final)
cat("Mediana:", round(mediana, 2), "\n")

# MODA (valor mas frecuente)
# Se crea una tabla de frecuencias y se busca el maximo
tabla_freq <- table(round(datos$Nota_Final, 0))
moda <- as.numeric(names(tabla_freq)[which.max(tabla_freq)])
cat("Moda:", moda, "\n")
```

Listing 9: Calcular media, mediana y moda de Nota Final

✎ Actividad

Se deben calcular las medidas de tendencia central para **Nota_Final** y anotar los resultados:

✍ Espacio para Respuesta

Media: _____

Mediana: _____

Moda: _____

6.2 Medidas de Dispersión

```r
# VARIANZA
varianza <- var(datos$Nota_Final)
cat("Varianza:", round(varianza, 2), "\n")

# DESVIACION ESTANDAR
desv_est <- sd(datos$Nota_Final)
cat("Desviacion Estandar:", round(desv_est, 2), "\n")

# CUARTILES
cuartiles <- quantile(datos$Nota_Final, probs = c(0.25, 0.50, 0.75))
cat("Q1 (25%):", round(cuartiles[1], 2), "\n")
cat("Q2 (50%):", round(cuartiles[2], 2), "\n")
cat("Q3 (75%):", round(cuartiles[3], 2), "\n")

# RANGO
rango <- max(datos$Nota_Final) - min(datos$Nota_Final)
cat("Rango:", round(rango, 2), "\n")
```

Listing 10: Calcular varianza, desviación estándar y cuartiles

✎ Actividad

Se deben calcular las medidas de dispersión para **Nota_Final**:

✍ Espacio para Respuesta

Varianza: _____ **Desviación Estándar:** _____

Q1: _____ **Q2:** _____ **Q3:** _____

Rango: _____

7 Visualización de Datos

7.1 Gráfico de Pastel: Categorías de Rendimiento

Primero, se crearán categorías basadas en la Nota Final.

```r
# Crear categorias de rendimiento
datos$Rendimiento <- cut(datos$Nota_Final,
  breaks = c(0, 60, 70, 80, 90, 100),
  labels = c("Reprobado", "Suficiente", "Bueno", "Muy Bueno",
     "Excelente"),
  include.lowest = TRUE
)

# Contar estudiantes por categoria
frecuencias <- table(datos$Rendimiento)
porcentajes <- round(prop.table(frecuencias) * 100, 1)

# Crear etiquetas con porcentajes
etiquetas <- paste(names(frecuencias), "\n", porcentajes, "%", sep = "")

# Definir colores
colores <- c("#FF6B6B", "#FFA94D", "#FFD93D", "#6BCB77", "#4D96FF")

# Crear grafico de pastel
pie(frecuencias,
    labels = etiquetas,
    col = colores,
    main = "Distribucion del Rendimiento Academico")
```

Listing 11: Crear categorías y gráfico de pastel

✎ Actividad

Se debe observar el gráfico de pastel y responder:

1. ¿Cuál es la categoría con más estudiantes?
2. ¿Qué porcentaje de estudiantes reprobó?

✍ Espacio para Respuesta

1. Categoría más frecuente: _____

2. Porcentaje de reprobados: _____

7.2 Diagrama de Barras Horizontales: Nivel de Asistencia

```r
# Crear categorias de asistencia
datos$Nivel_Asistencia <- cut(datos$Asistencia,
  breaks = c(0, 60, 75, 90, 100),
  labels = c("Bajo", "Regular", "Bueno", "Excelente"),
  include.lowest = TRUE
)

# Contar frecuencias
freq_asistencia <- table(datos$Nivel_Asistencia)

# Grafico de barras horizontales
barplot(freq_asistencia,
        horiz = TRUE,
        col = c("#FF6B6B", "#FFA94D", "#6BCB77", "#4D96FF"),
        main = "Estudiantes por Nivel de Asistencia",
        xlab = "Numero de Estudiantes",
        ylab = "Nivel de Asistencia",
        las = 1)
```

Listing 12: Crear categorías de asistencia y gráfico de barras

✎ Actividad

Se debe observar el gráfico de barras y responder:

1. ¿Cuántos estudiantes tienen asistencia “Excelente”?
2. ¿Cuál nivel de asistencia tiene menos estudiantes?

✍ Espacio para Respuesta

1. Estudiantes con asistencia Excelente: _____

2. Nivel con menos estudiantes: _____

7.3 Diagrama de Cajas: Exámenes Internos 1 y 2

✦ Concepto Teórico

El **diagrama de cajas (boxplot)** muestra:

- La **mediana** (línea central de la caja)
- Los **cuartiles Q1 y Q3** (bordes de la caja)
- Los **valores mínimo y máximo** (bigotes)
- Los **valores atípicos** (puntos fuera de los bigotes)

```r
# Crear diagrama de cajas para ambos examenes internos
boxplot(datos$Examen_Interno_1, datos$Examen_Interno_2,
        names = c("Examen Interno 1", "Examen Interno 2"),
        col = c("#4D96FF", "#6BCB77"),
        main = "Comparacion de Examenes Internos",
        ylab = "Calificacion (0-100)",
        border = "darkblue")

# Agregar linea de la media
abline(h = mean(datos$Examen_Interno_1), col = "red", lty = 2)
```

Listing 13: Diagrama de cajas comparando Examen Interno 1 y 2

```r
# Estadisticas del Examen Interno 1
cat("=== EXAMEN INTERNO 1 ===\n")
cat("Media:", round(mean(datos$Examen_Interno_1), 2), "\n")
cat("Mediana:", round(median(datos$Examen_Interno_1), 2), "\n")
cat("Desv. Est.:", round(sd(datos$Examen_Interno_1), 2), "\n\n")

# Estadisticas del Examen Interno 2
cat("=== EXAMEN INTERNO 2 ===\n")
cat("Media:", round(mean(datos$Examen_Interno_2), 2), "\n")
cat("Mediana:", round(median(datos$Examen_Interno_2), 2), "\n")
cat("Desv. Est.:", round(sd(datos$Examen_Interno_2), 2), "\n")
```

Listing 14: Estadísticas de ambos exámenes para comparar

✎ Actividad

Se deben comparar los dos exámenes internos usando el boxplot y las estadísticas:

1. ¿Cuál examen tiene la mediana más alta?
2. ¿Cuál examen tiene mayor dispersión (desviación estándar)?
3. ¿Se observan valores atípicos en alguno de los exámenes?

✍ Espacio para Respuesta

1. **Examen con mediana más alta:** _____

2. **Examen con mayor dispersión:** _____

3. **¿Existen valores atípicos?** _____

7.4 Gráfico de Violín: Nota Final por Nivel de Asistencia

✦ Concepto Teórico

El **gráfico de violín** combina un boxplot con una estimación de densidad. Muestra:

- La **forma de la distribución** (ancho del violín)
- La **mediana y cuartiles** (boxplot interior)
- La **densidad de los datos** en diferentes valores

Es útil para comparar distribuciones entre grupos.

```r
# Crear grafico de violin
ggplot(datos, aes(x = Nivel_Asistencia, y = Nota_Final,
                  fill = Nivel_Asistencia)) +
  geom_violin(trim = FALSE, alpha = 0.7) +
  geom_boxplot(width = 0.15, fill = "white") +
  scale_fill_manual(values = c("#FF6B6B", "#FFA94D",
                               "#6BCB77", "#4D96FF")) +
  labs(
    title = "Distribucion de Nota Final por Nivel de Asistencia",
    x = "Nivel de Asistencia",
    y = "Nota del Examen Final"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```

Listing 15: Gráfico de violín: Nota Final por Nivel de Asistencia

✎ Actividad

Se debe observar el gráfico de violín y responder:

1. ¿Cuál nivel de asistencia muestra las notas más altas?
2. ¿Se observa alguna diferencia en la forma de las distribuciones?

✍ Espacio para Respuesta

1. Nivel con notas más altas: _____

2. Diferencias en las distribuciones: _____

_____

8 Análisis de Normalidad: Nota del Examen Final

✦ Concepto Teórico

La **distribución normal** tiene forma de campana simétrica. Muchas pruebas estadísticas (como la correlación de Pearson) asumen que los datos son normales.
¿Cómo se evalúa la normalidad?

1. **Histograma:** Verificar si tiene forma de campana
2. **Q-Q Plot:** Los puntos deben seguir la línea diagonal
3. **Prueba de Shapiro-Wilk:** Si $p > 0.05$, los datos son normales
4. **Prueba de Kolmogorov-Smirnov:** Si $p > 0.05$, los datos son normales

8.1 Histograma de la Nota Final

```r
# Crear histograma
hist(datos$Nota_Final,
     breaks = 20,
     col = "#4D96FF",
     border = "white",
     main = "Histograma de Nota Final",
     xlab = "Nota del Examen Final",
     ylab = "Frecuencia",
     freq = FALSE)  # Usar densidad en lugar de frecuencia

# Agregar curva normal teorica
curve(dnorm(x, mean = mean(datos$Nota_Final), sd =
    sd(datos$Nota_Final)),
      add = TRUE,
      col = "red",
      lwd = 2)

# Agregar leyenda
legend("topright",
       legend = "Curva Normal",
       col = "red",
       lwd = 2)
```

Listing 16: Histograma con curva normal superpuesta

✎ Actividad

Se debe observar el histograma y responder:

1. ¿El histograma tiene forma de campana?
2. ¿La distribución parece simétrica o sesgada?

✍ Espacio para Respuesta

1. **¿Tiene forma de campana?** (Sí / No / Aproximadamente): _____

2. **¿Es simétrica o sesgada?** _____

8.2 Gráfico Q-Q (Quantile-Quantile)

```r
# Crear grafico Q-Q
qqnorm(datos$Nota_Final,
       main = "Grafico Q-Q: Nota Final",
       xlab = "Cuantiles Teoricos",
       ylab = "Cuantiles de los Datos",
       col = "#4D96FF",
       pch = 19,
       cex = 0.6)

# Agregar linea de referencia
qqline(datos$Nota_Final, col = "red", lwd = 2)
```

Listing 17: Gráfico Q-Q para evaluar normalidad

★ Nota Importante

En un Q-Q Plot:

- Si los puntos siguen la línea roja → los datos son normales
- Si los puntos se desvían en los extremos → existen colas pesadas o livianas
- Si los puntos forman una curva → los datos son asimétricos

✎ Actividad

Se debe observar el gráfico Q-Q y describir lo observado:

✍ Espacio para Respuesta

¿Los puntos siguen la línea roja? _____

¿Existen desviaciones en los extremos? _____

8.3 Pruebas Estadísticas de Normalidad

```r
# Prueba de Shapiro-Wilk
# NOTA: Esta prueba funciona mejor con menos de 5000 datos
shapiro_test <- shapiro.test(datos$Nota_Final)

cat("=== PRUEBA DE SHAPIRO-WILK ===\n")
cat("Estadistico W:", round(shapiro_test$statistic, 4), "\n")
cat("Valor p:", shapiro_test$p.value, "\n")

# Interpretar resultado
if(shapiro_test$p.value > 0.05) {
  cat("Conclusion: Los datos SON NORMALES (p > 0.05)\n")
} else {
  cat("Conclusion: Los datos NO SON NORMALES (p <= 0.05)\n")
```

```
}
```

Listing 18: Prueba de Shapiro-Wilk

```r
# Prueba de Kolmogorov-Smirnov
# Primero se estandarizan los datos
datos_std <- (datos$Nota_Final - mean(datos$Nota_Final)) /
    sd(datos$Nota_Final)
ks_test <- ks.test(datos_std, "pnorm")

cat("=== PRUEBA DE KOLMOGOROV-SMIRNOV ===\n")
cat("Estadistico D:", round(ks_test$statistic, 4), "\n")
cat("Valor p:", ks_test$p.value, "\n")

# Interpretar resultado
if(ks_test$p.value > 0.05) {
  cat("Conclusion: Los datos SON NORMALES (p > 0.05)\n")
} else {
  cat("Conclusion: Los datos NO SON NORMALES (p <= 0.05)\n")
}
```

Listing 19: Prueba de Kolmogorov-Smirnov

✎ Actividad

Se deben ejecutar las pruebas de normalidad y completar:

✍ Espacio para Respuesta

Prueba de Shapiro-Wilk:

Valor p = _____ Conclusión: _____

Prueba de Kolmogorov-Smirnov:

Valor p = _____ Conclusión: _____

¿Los datos de Nota_Final son normales? _____

9 Visualizaciones Avanzadas

9.1 Mapa de Calor (Heatmap) de Correlaciones

✦ Concepto Teórico

El **mapa de calor** permite visualizar la matriz de correlaciones entre todas las variables numéricas. Los colores indican la fuerza y dirección de la correlación:

- **Colores cálidos (rojos):** Correlación positiva
- **Colores fríos (azules):** Correlación negativa
- **Colores neutros (blancos):** Correlación cercana a cero

```r
# Seleccionar solo las variables numericas (sin ID)
vars_numericas <- datos[, c("Asistencia", "Examen_Interno_1",
                            "Examen_Interno_2", "Promedio_Tareas",
                            "Horas_Estudio", "Nota_Final")]

# Calcular la matriz de correlacion
matriz_cor <- cor(vars_numericas)

# Ver la matriz de correlacion
round(matriz_cor, 2)
```

Listing 20: Crear matriz de correlación

```r
# Crear mapa de calor basico
heatmap(matriz_cor,
        col = colorRampPalette(c("#4D96FF", "white", "#FF6B6B"))(20),
        symm = TRUE,
        main = "Mapa de Calor - Correlaciones")
```

Listing 21: Crear mapa de calor con heatmap básico

```r
# Instalar si es necesario: install.packages("reshape2")
library(reshape2)

# Convertir matriz a formato largo
matriz_largo <- melt(matriz_cor)
colnames(matriz_largo) <- c("Variable1", "Variable2", "Correlacion")

# Crear mapa de calor con ggplot2
ggplot(matriz_largo, aes(x = Variable1, y = Variable2, fill =
    Correlacion)) +
  geom_tile(color = "white") +
  geom_text(aes(label = round(Correlacion, 2)), size = 3) +
  scale_fill_gradient2(low = "#4D96FF", mid = "white", high = "#FF6B6B",
                       midpoint = 0, limit = c(-1, 1)) +
  labs(title = "Mapa de Calor - Matriz de Correlaciones",
       x = "", y = "") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Listing 22: Mapa de calor mejorado con ggplot2

✎ Actividad

Se debe observar el mapa de calor y responder:

1. ¿Cuáles dos variables tienen la correlación más fuerte (más cercana a 1)?
2. ¿Existe alguna correlación negativa entre las variables?
3. ¿Cuál variable tiene la mayor correlación con Nota_Final?

✍ Espacio para Respuesta

1. Variables con correlación más fuerte: _____

2. ¿Existe correlación negativa? _____

3. Variable más correlacionada con Nota_Final: _____

9.2 Gráfico de Violín

✦ Concepto Teórico

El **gráfico de violín** combina un diagrama de cajas con una estimación de densidad. Permite visualizar:

- La **distribución completa** de los datos (forma del violín)
- La **mediana y cuartiles** (caja interior)
- La **concentración** de datos en diferentes valores

Es útil para comparar distribuciones entre grupos.

```r
# Crear grafico de violin
ggplot(datos, aes(x = Nivel_Asistencia, y = Nota_Final,
                  fill = Nivel_Asistencia)) +
  geom_violin(trim = FALSE, alpha = 0.7) +
  geom_boxplot(width = 0.15, fill = "white") +
  scale_fill_manual(values = c("#FF6B6B", "#FFA94D", "#6BCB77",
     "#4D96FF")) +
  labs(title = "Distribucion de Nota Final por Nivel de Asistencia",
       x = "Nivel de Asistencia",
       y = "Nota del Examen Final") +
  theme_minimal() +
  theme(legend.position = "none")
```

Listing 23: Gráfico de violín: Nota Final por Nivel de Asistencia

```r
# Grafico de violin por categoria de rendimiento
ggplot(datos, aes(x = Rendimiento, y = Examen_Interno_1,
                  fill = Rendimiento)) +
  geom_violin(trim = FALSE, alpha = 0.7) +
  geom_boxplot(width = 0.1, fill = "white") +
  scale_fill_manual(values = c("#FF6B6B", "#FFA94D", "#FFD93D",
                               "#6BCB77", "#4D96FF")) +
  labs(title = "Distribucion de Examen Interno 1 por Rendimiento Final",
       x = "Categoria de Rendimiento",
       y = "Nota Examen Interno 1") +
  theme_minimal() +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 45, hjust = 1))
```

Listing 24: Gráfico de violín: Nota Final por Rendimiento

✎ Actividad

Se debe observar el gráfico de violín de Nota Final por Nivel de Asistencia y responder:

1. ¿Cuál nivel de asistencia muestra la distribución más concentrada (violín más angosto)?
2. ¿En cuál nivel de asistencia se observan las notas más altas?
3. ¿Qué se puede concluir sobre la relación entre asistencia y rendimiento?

✍ Espacio para Respuesta

1. Nivel con distribución más concentrada: _____

2. Nivel con notas más altas: _____

3. Conclusión sobre asistencia y rendimiento: _____

_____

10 Análisis de Correlación

Primero se analizará la correlación entre todas las variables numéricas mediante un mapa de calor, y luego se profundizará en la relación entre **Asistencia** y **Nota Final**.

✦ Concepto Teórico

El **coeficiente de correlación de Pearson** (r) mide la fuerza y dirección de la relación lineal entre dos variables.
Interpretación:

- $r = 1$: Correlación positiva perfecta
- $0.7 \leq r < 1$: Correlación positiva fuerte
- $0.4 \leq r < 0.7$: Correlación positiva moderada
- $0.2 \leq r < 0.4$: Correlación positiva débil
- $-0.2 < r < 0.2$: Correlación muy débil o nula
- $r = -1$: Correlación negativa perfecta

10.1 Mapa de Calor (Heatmap) de Correlaciones

```r
# Seleccionar solo variables numericas
vars_numericas <- datos[, c("Asistencia", "Examen_Interno_1",
                            "Examen_Interno_2", "Promedio_Tareas",
                            "Horas_Estudio", "Nota_Final")]

# Calcular matriz de correlacion
matriz_cor <- cor(vars_numericas)

# Mostrar la matriz
round(matriz_cor, 2)
```

Listing 25: Calcular matriz de correlación

```r
# Mapa de calor con corrplot
corrplot(matriz_cor,
         method = "color",
         type = "upper",
         addCoef.col = "black",
         number.cex = 0.7,
         tl.col = "black",
         tl.srt = 45,
         col = colorRampPalette(c("#FF6B6B", "white", "#4D96FF"))(100),
         title = "Mapa de Calor - Correlaciones",
         mar = c(0, 0, 2, 0))
```

Listing 26: Crear mapa de calor con corrplot

```r
# Convertir matriz a formato largo
matriz_largo <- melt(matriz_cor)
```

```r
colnames(matriz_largo) <- c("Variable1", "Variable2", "Correlacion")

# Crear heatmap con ggplot2
ggplot(matriz_largo, aes(x = Variable1, y = Variable2,
                         fill = Correlacion)) +
  geom_tile(color = "white") +
  geom_text(aes(label = round(Correlacion, 2)), size = 3) +
  scale_fill_gradient2(low = "#FF6B6B", mid = "white",
                       high = "#4D96FF", midpoint = 0) +
  labs(title = "Matriz de Correlacion") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Listing 27: Mapa de calor alternativo con ggplot2

✎ Actividad

Se debe observar el mapa de calor y responder:

1. ¿Cuáles dos variables tienen la correlación más alta?
2. ¿Existe alguna correlación negativa?
3. ¿Qué variable tiene la correlación más fuerte con Nota_Final?

✍ Espacio para Respuesta

1. Variables con correlación más alta: _____

2. ¿Existe correlación negativa? _____

3. Variable más correlacionada con Nota_Final: _____

10.2 Gráfico de Dispersión: Asistencia vs Nota Final

```r
# Crear grafico de dispersion
plot(datos$Asistencia, datos$Nota_Final,
     main = "Relacion entre Asistencia y Nota Final",
     xlab = "Porcentaje de Asistencia",
     ylab = "Nota del Examen Final",
     col = "#4D96FF",
     pch = 19,
     cex = 0.7)

# Agregar linea de tendencia
abline(lm(Nota_Final ~ Asistencia, data = datos),
       col = "red",
       lwd = 2)
```

Listing 28: Gráfico de dispersión: Asistencia vs Nota Final

10.3 Cálculo del Coeficiente de Correlación

```r
# Calcular correlacion de Pearson
correlacion <- cor(datos$Asistencia, datos$Nota_Final)
cat("Coeficiente de correlacion (r):", round(correlacion, 4), "\n")

# Prueba de significancia
prueba_cor <- cor.test(datos$Asistencia, datos$Nota_Final)
cat("Valor p:", prueba_cor$p.value, "\n")

# Interpretar
if(prueba_cor$p.value < 0.05) {
  cat("La correlacion ES estadisticamente significativa\n")
} else {
  cat("La correlacion NO es significativa\n")
}
```

Listing 29: Calcular correlación de Pearson

✎ Actividad

Se debe calcular la correlación entre Asistencia y Nota Final:

✍ **Espacio para Respuesta**

Coeficiente r = _____

Valor p = _____

Tipo de correlación (fuerte/moderada/débil): _____

¿Es significativa? (Sí / No): _____

11 Regresión Lineal Simple

Se creará un modelo para predecir la **Nota Final** basándose en la **Asistencia**.

✦ Concepto Teórico

La **regresión lineal simple** modela la relación entre dos variables con la ecuación:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Donde:

- Y: Variable dependiente (Nota Final)
- X: Variable independiente (Asistencia)
- β_0: Intercepto (valor de Y cuando X = 0)
- β_1: Pendiente (cambio en Y por cada unidad de X)
- ε: Error aleatorio

El **coeficiente de determinación** (R^2) indica qué porcentaje de la variabilidad de Y es explicado por X.

11.1 Crear el Modelo de Regresión

```r
# Crear el modelo de regresion
modelo <- lm(Nota_Final ~ Asistencia, data = datos)

# Ver resumen del modelo
summary(modelo)
```

Listing 30: Crear modelo de regresión lineal simple

11.2 Interpretar los Resultados

```r
# Obtener coeficientes
intercepto <- coef(modelo)[1]
pendiente <- coef(modelo)[2]

cat("=== ECUACION DEL MODELO ===\n")
cat("Nota_Final =", round(intercepto, 2), "+",
    round(pendiente, 2), "* Asistencia\n\n")

# Coeficiente de determinacion R^2
r_cuadrado <- summary(modelo)$r.squared
cat("R cuadrado (R^2):", round(r_cuadrado, 4), "\n")
cat("Esto significa que la Asistencia explica el",
    round(r_cuadrado * 100, 1), "% de la variabilidad en la Nota
       Final\n")
```

Listing 31: Extraer e interpretar coeficientes

✎ Actividad

Se debe crear el modelo de regresión y completar:

✍ Espacio para Respuesta

Ecuación del modelo:

Nota_Final = _____ + _____ × Asistencia

R cuadrado (R^2) = _____

Interpretación: La Asistencia explica el _____% de la variabilidad en la nota final.

11.3 Visualizar el Modelo de Regresión

```r
# Grafico con ggplot2
ggplot(datos, aes(x = Asistencia, y = Nota_Final)) +
  geom_point(color = "#4D96FF", alpha = 0.5) +
  geom_smooth(method = "lm", color = "red", se = TRUE) +
  labs(
    title = "Regresion Lineal: Nota Final vs Asistencia",
    subtitle = paste("R^2 =", round(r_cuadrado, 4)),
    x = "Porcentaje de Asistencia",
    y = "Nota del Examen Final"
  ) +
  theme_minimal()
```

Listing 32: Gráfico del modelo de regresión

11.4 Realizar Predicciones

```r
# Crear datos para prediccion
nuevos_datos <- data.frame(
  Asistencia = c(60, 70, 80, 90, 100)
)

# Realizar predicciones
predicciones <- predict(modelo, newdata = nuevos_datos)

# Mostrar resultados
cat("=== PREDICCIONES ===\n")
for(i in 1:5) {
  cat("Si la asistencia es", nuevos_datos$Asistencia[i],
      "% -> Nota predicha:",
      round(predicciones[i], 1), "\n")
}
```

Listing 33: Predecir notas para diferentes niveles de asistencia

✎ Actividad

Se debe usar el modelo para predecir la nota de un estudiante que tiene **85% de asistencia**:

```r
# Predecir para 85% de asistencia
mi_prediccion <- data.frame(Asistencia = 85)
resultado <- predict(modelo, newdata = mi_prediccion)
cat("Nota predicha para 85% de asistencia:", round(resultado, 1), "\n")
```

Listing 34: Predicción personalizada

✍ Espacio para Respuesta

Si un estudiante tiene 85% de asistencia:

Nota Final predicha = _____

12 Actividades de Cierre

✎ Actividad

Reflexión Final
Con base en todo el análisis realizado, se deben responder las siguientes preguntas:

1. ¿Cuál es el promedio de la Nota Final de todos los estudiantes?
2. ¿Los datos de Nota Final siguen una distribución normal? Se debe justificar la respuesta.
3. ¿Existe una relación significativa entre la Asistencia y la Nota Final?
4. Según el modelo, ¿cuánto aumenta la nota final por cada punto porcentual adicional de asistencia?
5. ¿Qué recomendaciones se le darían a un estudiante que desea mejorar su nota final?

✍ Espacio para Respuesta

1. Promedio de Nota Final: _____

2. ¿Son normales los datos? _____

Justificación: _____

_____

3. ¿Existe relación significativa? _____

4. Aumento por punto de asistencia: _____

5. Recomendaciones: _____

_____

_____

13 Resumen del Taller

Funciones de R aprendidas

1. **Importación de datos:** `read.csv()`
2. **Exploración:** `head()`, `str()`, `summary()`
3. **Estadística descriptiva:** `mean()`, `median()`, `sd()`, `var()`, `quantile()`
4. **Visualización:** `pie()`, `barplot()`, `boxplot()`, `hist()`
5. **Normalidad:** `qqnorm()`, `qqline()`, `shapiro.test()`, `ks.test()`
6. **Correlación:** `cor()`, `cor.test()`
7. **Regresión:** `lm()`, `predict()`

14 Referencias y Recursos

14.1 Dataset

- Fuente: Kaggle - Student Performance Prediction Dataset
- URL: `https://www.kaggle.com/`

14.2 Recursos de R

- R Documentation: `https://www.rdocumentation.org/`
- R for Data Science: `https://r4ds.had.co.nz/`
- ggplot2: `https://ggplot2.tidyverse.org/`

Colegio Universitario de Cartago
Profesor: David Martínez Salazar
III Cuatrimestre 2025
Taller elaborado con fines educativos