TAREA 1

María Pallares Diez

2025-01-05

Introducción

En este informe, se analiza el conjunto de datos "Rendimiento Académico y Hábitos". El análisis incluye un estudio univariante, la detección de valores atípicos, las relaciones entre variables y la comparación de estudiantes promedio por centro educativo.

Carga de Datos

Se carga el archivo CSV y se verifica su estructura.

```
data <- read.csv("BD_Rendimiento_Academico.csv", sep = ";", encoding = "latin1")
head(data)
```

```
##   ID Edad Horas_estudio Promedio_matematicas Promedio_ciencias Promedio_lectura
## 1  1 17.0           7.9                 78.6              66.7             50.9
## 2  2 16.4          12.9                 80.6              69.4             62.8
## 3  3 17.1          13.3                 85.8              82.5             70.1
## 4  4 18.0          11.0                 85.5              81.1             70.6
## 5  5 16.3          14.2                 61.2              74.8             64.6
## 6  6 16.3          17.0                 65.6              76.2             77.5
##   Asistencia Horas_sueño Nivel_estres Uso_dispositivos Condicion_fisica Centro
## 1       85.8         7.2          2.6              4.1            184.1      6
## 2       82.8         7.2          6.4              5.2            242.3      6
## 3       74.2         6.5          4.5              4.7            179.2      9
## 4       99.5         6.2          4.3              3.9            132.0      9
## 5       84.2         7.8          6.4              2.9            179.5      2
## 6       69.5         6.1          5.9              0.0            205.4      3
```

Análisis Univariante

Se presentan resúmenes descriptivos y gráficos para cada variable numérica del conjunto de datos.

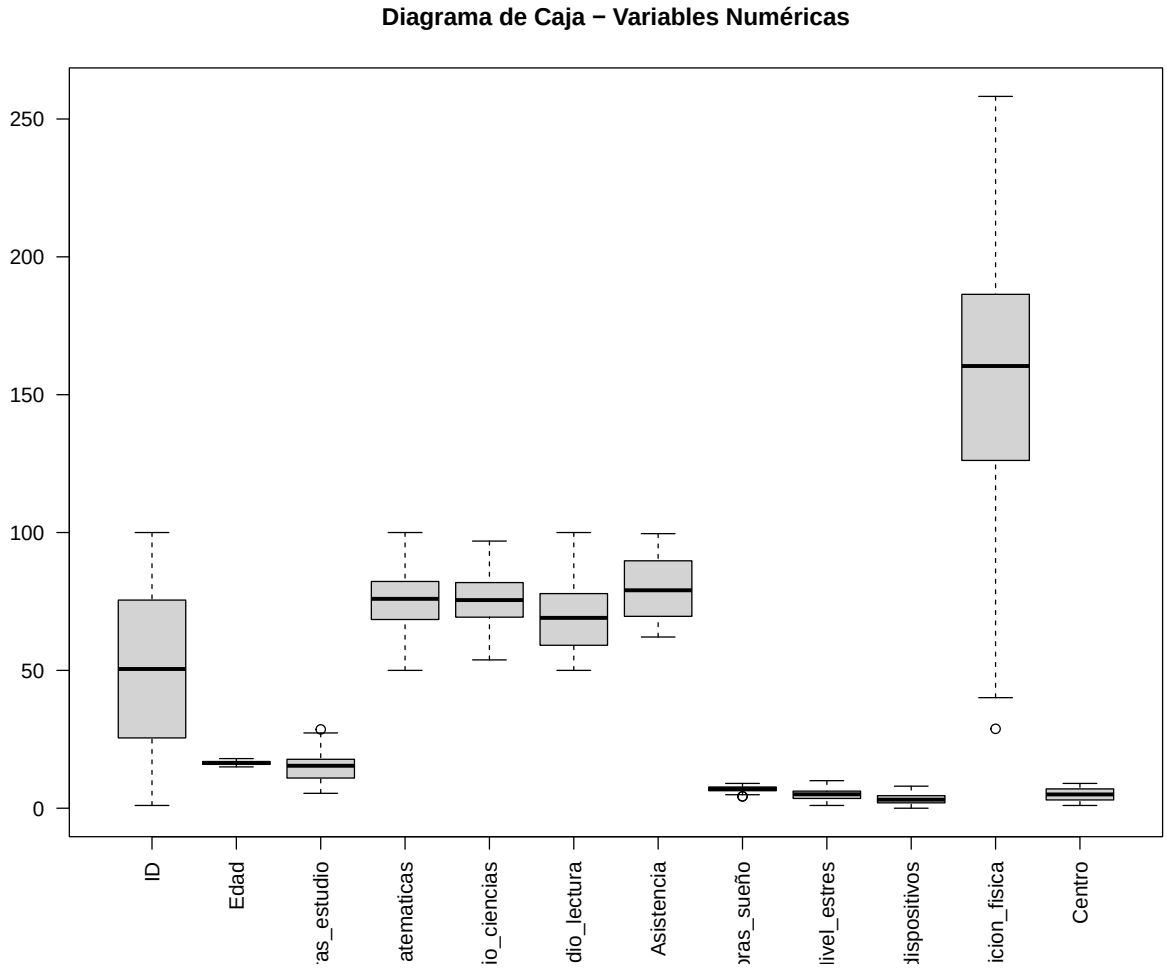
```
summary(data)
```

```
##        ID              Edad       Horas_estudio   Promedio_matematicas
##  Min.   :  1.00   Min.   :15.00   Min.   : 5.40   Min.   : 50.00
##  1st Qu.: 25.75   1st Qu.:15.90   1st Qu.:10.97   1st Qu.: 68.47
##  Median : 50.50   Median :16.40   Median :15.40   Median : 75.95
##  Mean   : 50.50   Mean   :16.42   Mean   :15.11   Mean   : 75.59
```

```
##  3rd Qu.: 75.25   3rd Qu.:16.93   3rd Qu.:17.68   3rd Qu.: 82.08
##  Max.   :100.00   Max.   :18.00   Max.   :28.60   Max.   :100.00
##  Promedio_ciencias Promedio_lectura   Asistencia     Horas_sueño
##  Min.   :53.80     Min.   : 50.00   Min.   :62.10   Min.   :4.300
##  1st Qu.:69.35     1st Qu.: 59.30   1st Qu.:69.65   1st Qu.:6.375
##  Median :75.50     Median : 69.05   Median :79.05   Median :6.950
##  Mean   :76.07     Mean   : 69.44   Mean   :79.61   Mean   :6.997
##  3rd Qu.:81.83     3rd Qu.: 77.78   3rd Qu.:89.67   3rd Qu.:7.700
##  Max.   :96.90     Max.   :100.00   Max.   :99.60   Max.   :9.000
##   Nivel_estres    Uso_dispositivos Condicion_fisica     Centro
##  Min.   : 1.000   Min.   :0.000    Min.   : 28.8    Min.   :1.00
##  1st Qu.: 3.575   1st Qu.:1.975    1st Qu.:126.8    1st Qu.:3.00
##  Median : 5.050   Median :3.100    Median :160.4    Median :5.00
##  Mean   : 4.960   Mean   :3.283    Mean   :156.9    Mean   :4.98
##  3rd Qu.: 6.200   3rd Qu.:4.475    3rd Qu.:185.9    3rd Qu.:7.00
##  Max.   :10.000   Max.   :8.000    Max.   :258.2    Max.   :9.00
```

Interpretación: Los estadísticos básicos (media, mediana, desviación estándar, etc.) nos permiten entender la distribución de las variables. Por ejemplo, la media de horas de estudio semanal es de 15.11, indicando que los estudiantes dedican en promedio 15 horas semanales al estudio.

```
boxplot(data[, sapply(data, is.numeric)], main = "Diagrama de Caja - Variables Numéricas", las = 2)
```



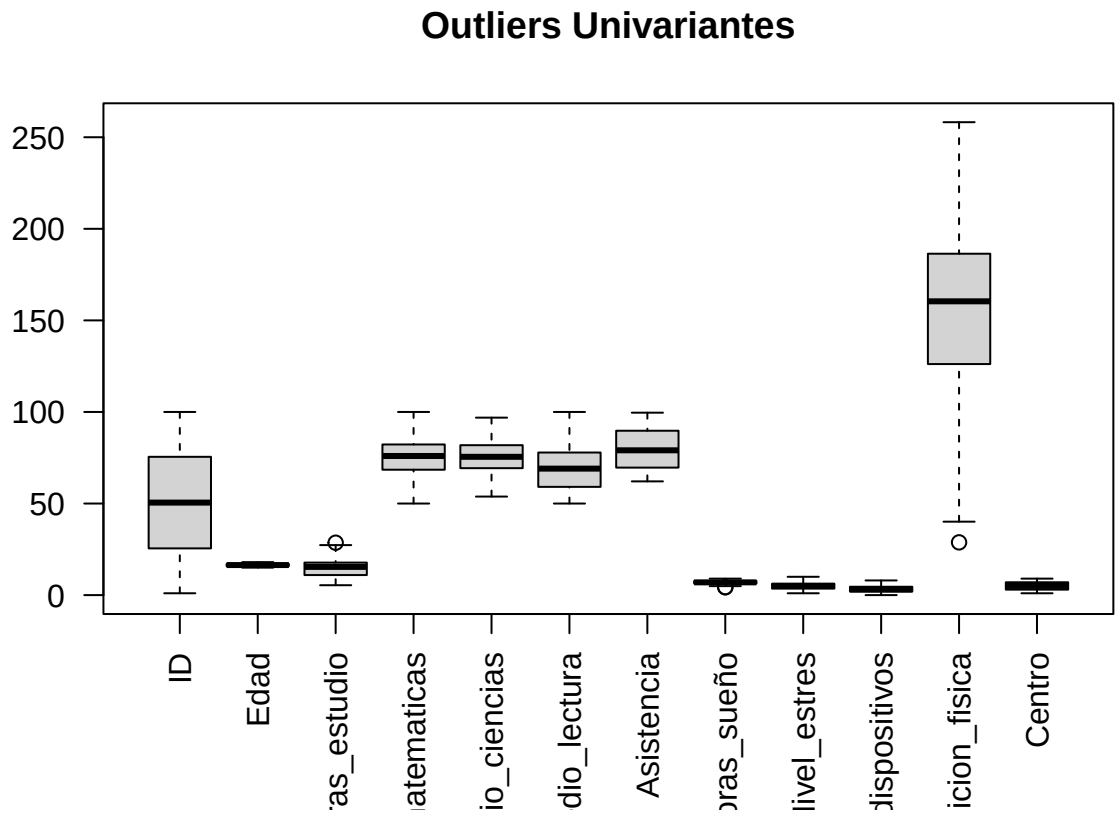
Interpretación: El diagrama de caja destaca posibles valores atípicos en variables como “Horas de estudio” y “Condición física”.

Detección de Outliers

Se identifican valores atípicos univariantes y multivariantes.

Univariante

```
boxplot(data[, sapply(data, is.numeric)], main = "Outliers Univariantes", las = 2)
```



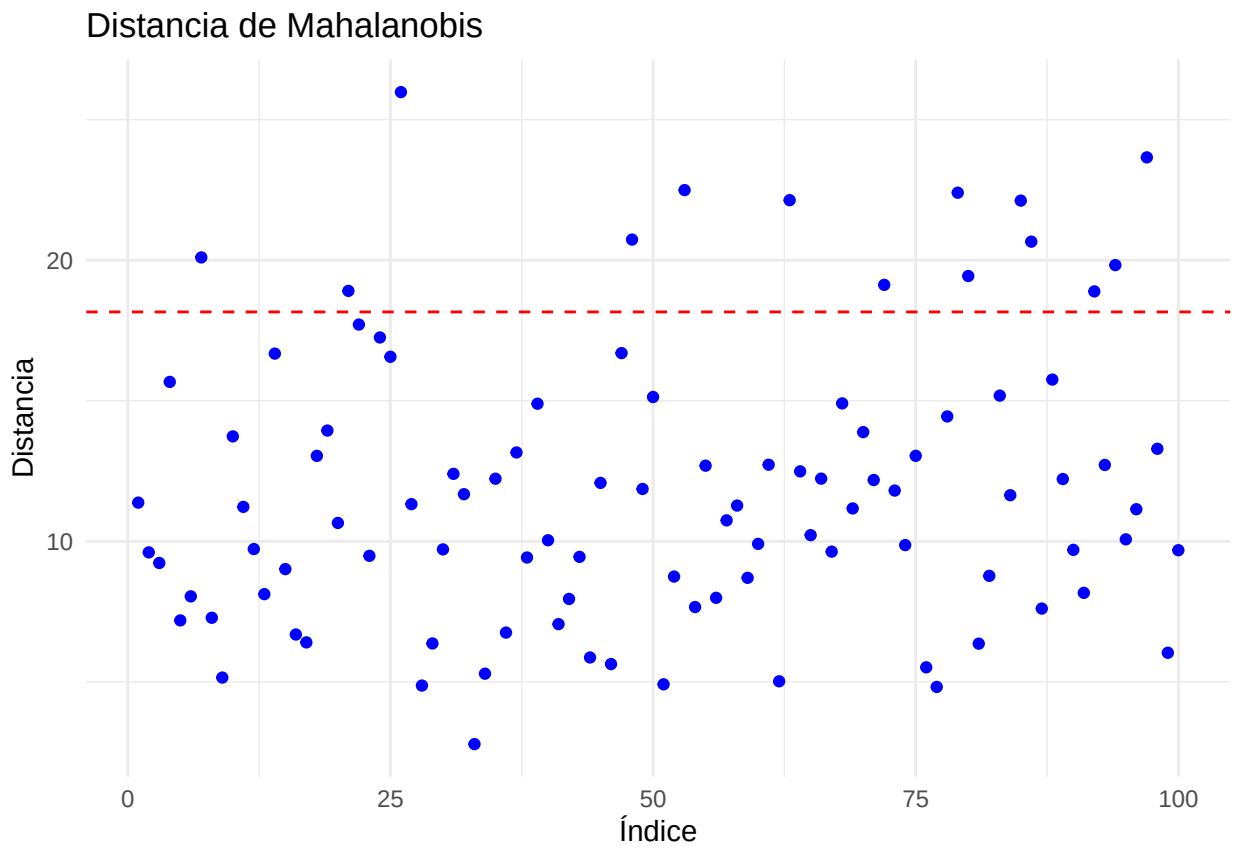
Interpretación: Se observan valores atípicos extremos en las variables de asistencia y condición física.

Multivariante

Se calcula la distancia de Mahalanobis.

```
# Seleccionar columnas numéricas
numeric_data <- data[, sapply(data, is.numeric)]

# Manejar valores NA, si existen
numeric_data <- na.omit(numeric_data)
```

```r
# Calcular la distancia de Mahalanobis
mahal_dist <- mahalanobis(
  numeric_data,
  colMeans(numeric_data, na.rm = TRUE),
  cov(numeric_data)
)

# Umbral para detección de outliers
threshold <- sqrt(ncol(numeric_data)) + 3 * sqrt(2 * ncol(numeric_data))

# Crear un gráfico usando ggplot2
library(ggplot2)
plot_data <- data.frame(Index = 1:length(mahal_dist), Distance = mahal_dist)

# Generar el gráfico
ggplot(plot_data, aes(x = Index, y = Distance)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = threshold, color = "red", linetype = "dashed") +
  labs(
    title = "Distancia de Mahalanobis",
    x = "Índice",
    y = "Distancia"
  ) +
  theme_minimal()
```



Interpretación: Los puntos que superan el umbral rojo representan posibles outliers multivariantes.

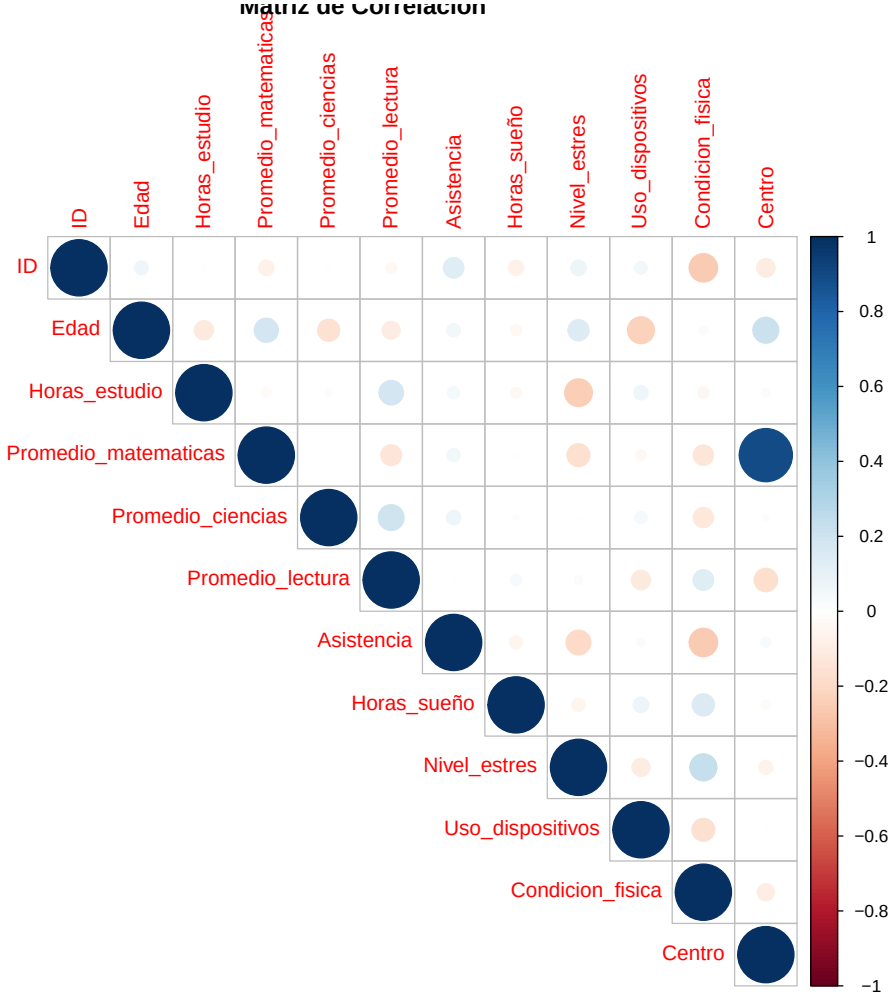
Relación entre Variables

Se exploran las relaciones utilizando una matriz de correlación.

```
cor_matrix <- cor(data[, sapply(data, is.numeric)], use = "complete.obs")
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
corrplot(cor_matrix, method = "circle", type = "upper", title = "Matriz de Correlación")
```

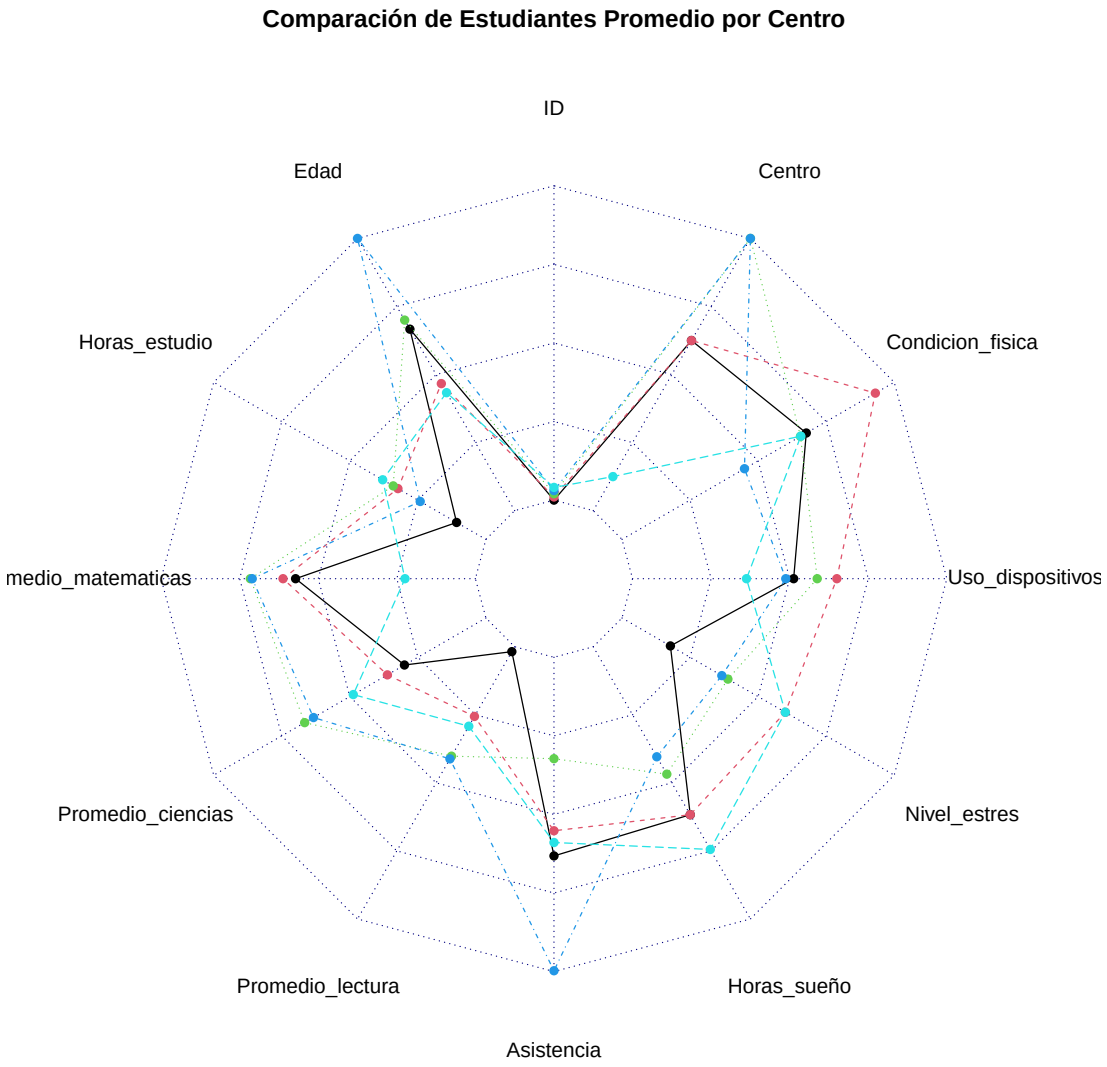


Interpretación: Las variables "Promedio Matemáticas" y "Horas de estudio" tienen una correlación positiva significativa, sugiriendo que mayores horas de estudio están asociadas con mejor rendimiento en matemáticas.

Comparativa entre Individuos

Se utiliza un gráfico de radar para comparar estudiantes promedio por centro.

```r
library(fmsb)
normalized_data <- as.data.frame(lapply(data[, sapply(data, is.numeric)],
                                        function(x) (x - min(x)) / (max(x) - min(x))))
max_min <- rbind(apply(normalized_data, 2, max), apply(normalized_data, 2, min))
radarchart(rbind(max_min, normalized_data[1:5, ]),
           title = "Comparación de Estudiantes Promedio por Centro")
```



Interpretación: El gráfico de radar muestra diferencias significativas en el rendimiento académico y hábitos de los estudiantes entre centros educativos.

Conclusiones

El análisis exploratorio nos permitió: 1. Identificar valores atípicos en variables clave. 2. Explorar correlaciones significativas entre variables académicas y de estilo de vida. 3. Comparar patrones entre centros educativos.

Este análisis proporciona una base para estudios más avanzados y estrategias de mejora académica.