Tarea - Reducción de dimensión para la base de datos habits.csv

Equipo n

2025-08-20

Análisis Exploratoio de Datos

El conjunto de datos de `habits.csv` contiene 1000 observaciones de estudiantes y 15 variables relacionadas con sus hábitos y rendimiento académico. La información incluye datos demográficos, hábitos de estudio y estilo de vida, y una calificación final en los exámenes.

La información incluye datos sobre hábitos de estudio y estilo de vida, y una calificación final en los exámenes que es la variable objetivo.

Como podemos notar de la descripción de los datos, tenemos dos tipos de variables en cada uno de los datos.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
habits |>
  select(age, study_hours_per_day, social_media_hours, netflix_hours, attendance_percentage, sleep_hours
  summary()
```

```
##       age        study_hours_per_day social_media_hours netflix_hours
##  Min.   :17.00   Min.   :0.00        Min.   :0.000      Min.   :0.000
##  1st Qu.:18.75   1st Qu.:2.60        1st Qu.:1.700      1st Qu.:1.000
##  Median :20.00   Median :3.50        Median :2.500      Median :1.800
##  Mean   :20.50   Mean   :3.55        Mean   :2.506      Mean   :1.820
##  3rd Qu.:23.00   3rd Qu.:4.50        3rd Qu.:3.300      3rd Qu.:2.525
```

```
##  Max.   :24.00   Max.   :8.30        Max.   :7.200      Max.   :5.400
##  attendance_percentage  sleep_hours    exercise_frequency mental_health_rating
##  Min.   : 56.00        Min.   : 3.20   Min.   :0.000      Min.   : 1.000
##  1st Qu.: 78.00        1st Qu.: 5.60   1st Qu.:1.000      1st Qu.: 3.000
##  Median : 84.40        Median : 6.50   Median :3.000      Median : 5.000
##  Mean   : 84.13        Mean   : 6.47   Mean   :3.042      Mean   : 5.438
##  3rd Qu.: 91.03        3rd Qu.: 7.30   3rd Qu.:5.000      3rd Qu.: 8.000
##  Max.   :100.00        Max.   :10.00   Max.   :6.000      Max.   :10.000
##    exam_score
##  Min.   : 18.40
##  1st Qu.: 58.48
##  Median : 70.50
##  Mean   : 69.60
##  3rd Qu.: 81.33
##  Max.   :100.00
```

```
habits %>%
  select(gender, part_time_job, diet_quality, parental_education_level, internet_quality, extracurricula
  lapply(table)
```

```
## $gender
##
## Female   Male  Other
##    481    477     42
##
## $part_time_job
##
##  No Yes
## 785 215
##
## $diet_quality
##
## Fair Good Poor
##  437  378  185
##
## $parental_education_level
##
##    Bachelor High School      Master        None
##         350         392         167          91
##
## $internet_quality
##
## Average    Good    Poor
##     391     447     162
##
## $extracurricular_participation
##
##  No Yes
## 682 318
```

Estadísticos descriptivos.

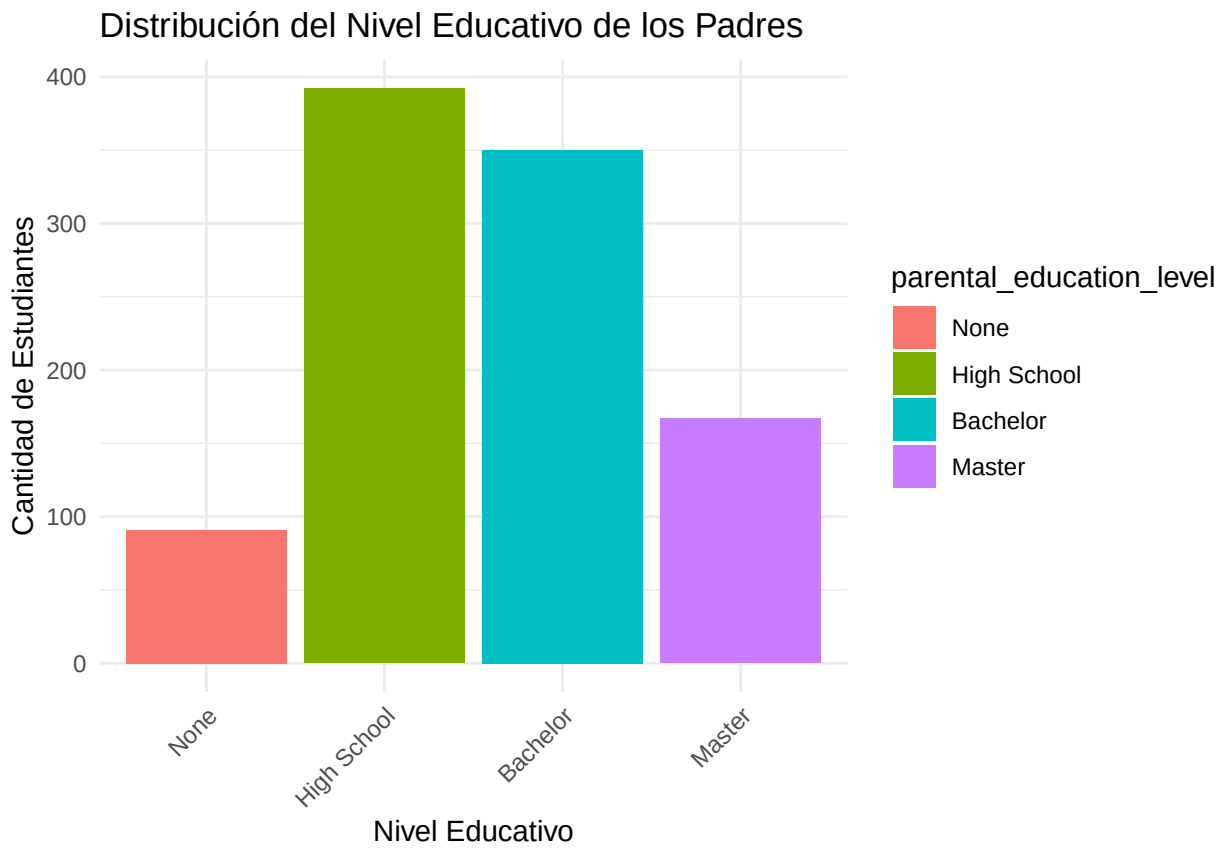
Variables numéricas

- **age**: La edad de los estudiantes, con un rango de 17 a 24 años. La media es de aproximadamente 20.5 años.
- **study_hours_per_day**: Horas de estudio diarias. El promedio es de 3.55 horas. El rango va de 0 a 8.3 horas.
- **social_media_hours**: Horas de uso de redes sociales al día. El promedio es de 2.5 horas, con un rango de 0 a 5.7 horas.
- **netflix_hours**: Horas de Netflix al día. El promedio es de 1.82 horas. El rango es de 0 a 5.4 horas.
- **attendance_percentage**: Porcentaje de asistencia a clases. El promedio es de 84.13%, con un rango de 56.1% a 100%.
- **sleep_hours**: Horas de sueño al día. El promedio es de 6.47 horas. El rango va de 3.20 a 10 horas.
- **exercise_frequency**: Frecuencia de ejercicio por semana. El promedio es de 3.04 veces. El rango va de 0 a 6 veces.
- **mental_health_rating**: Calificación de salud mental. El promedio es de 5.5, con un rango de 1 a 10.
- **exam_score**: La variable objetivo, el puntaje del examen. El promedio es de 67.9 puntos. El rango va de 18.4 a 100, lo que indica un amplio rango de rendimiento académico.

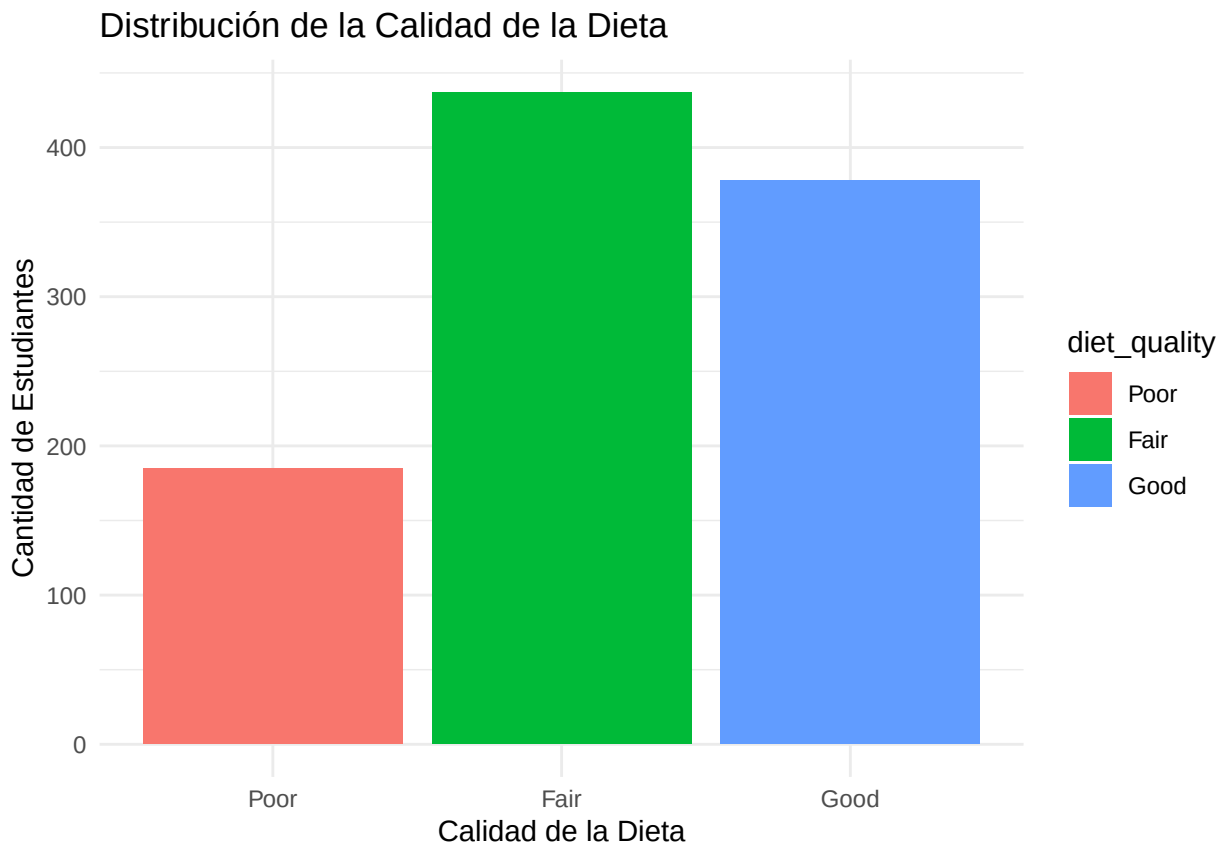
Para una mejor visualización, podemos convertir algunas variables de texto a "factores" y reordenarlas de forma lógica.

```
datos <- habits |>
  mutate(
    diet_quality = factor(diet_quality, levels = c("Poor", "Fair", "Good")),
    internet_quality = factor(internet_quality, levels = c("Poor", "Average", "Good")),
    parental_education_level = factor(parental_education_level,
                                      levels = c("None", "High School", "Bachelor", "Master"))
  )
```

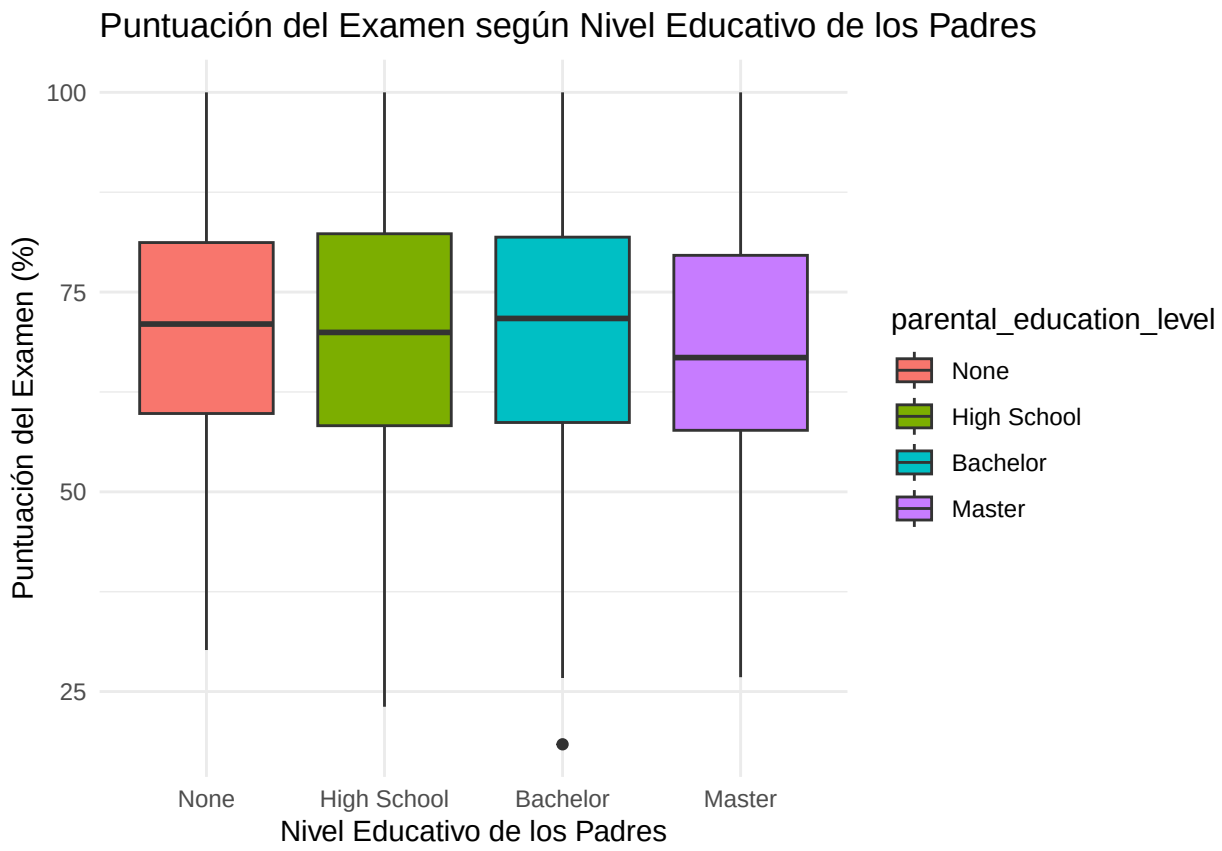
```
ggplot(datos, aes(x = parental_education_level, fill = parental_education_level)) +
  geom_bar() +
  labs(title = "Distribución del Nivel Educativo de los Padres",
       x = "Nivel Educativo",
       y = "Cantidad de Estudiantes") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



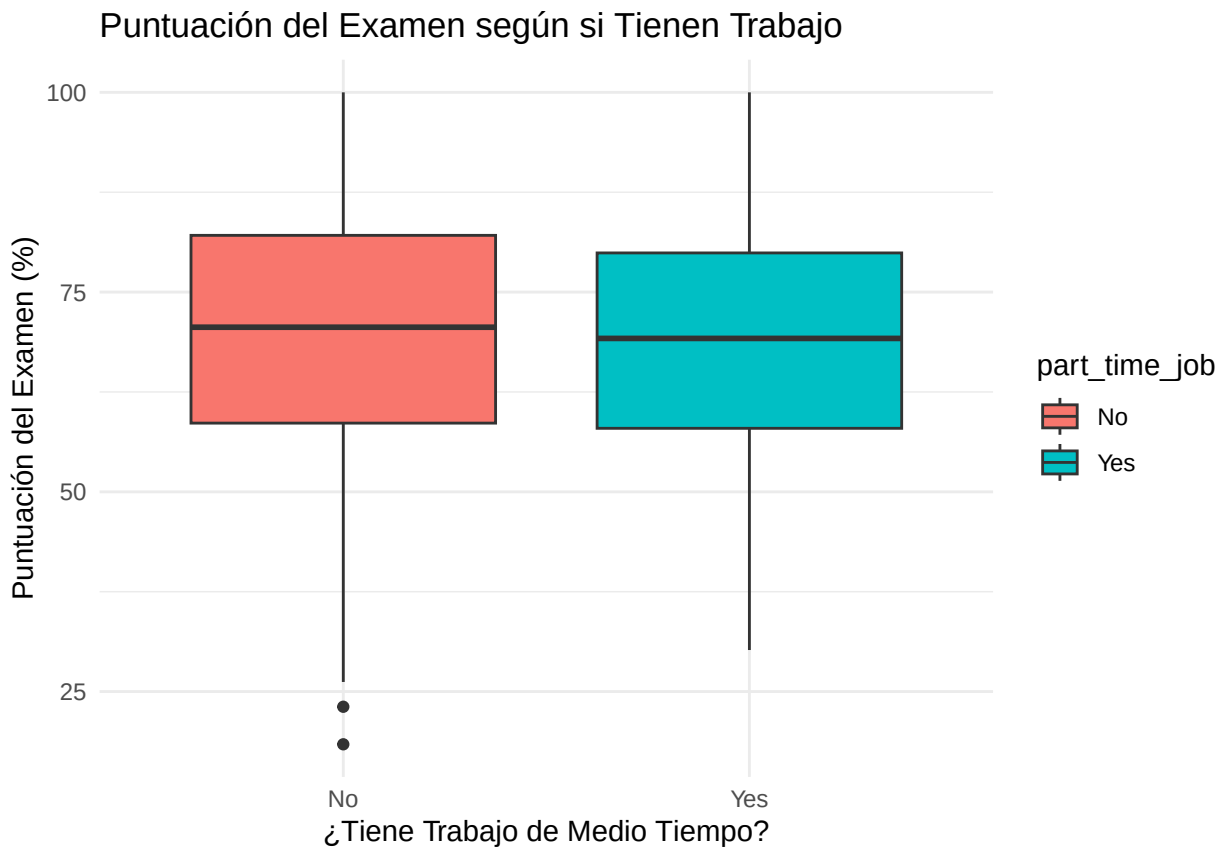

```r
# Gráfico 2: Distribución de la Calidad de la Dieta
ggplot(datos, aes(x = diet_quality, fill = diet_quality)) +
  geom_bar() +
  labs(title = "Distribución de la Calidad de la Dieta",
       x = "Calidad de la Dieta",
       y = "Cantidad de Estudiantes") +
  theme_minimal()
```



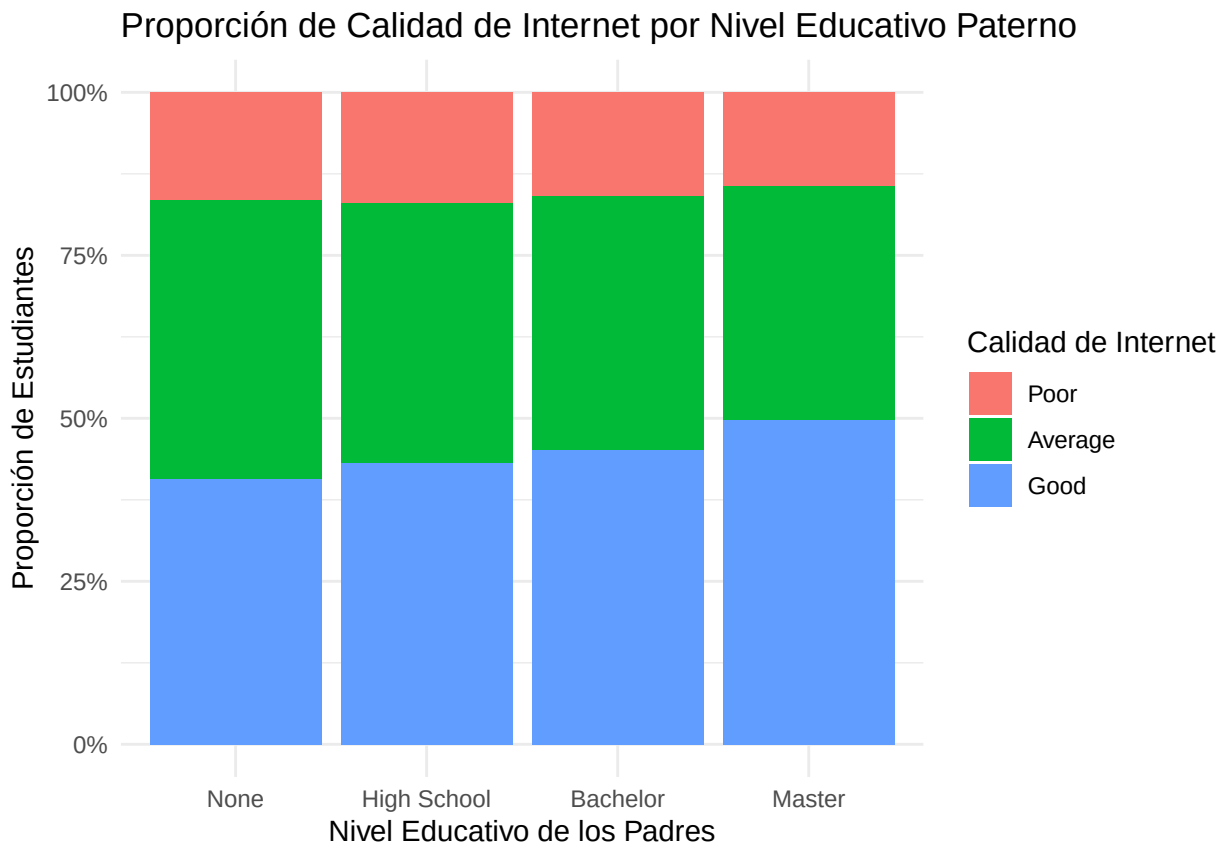

```r
ggplot(datos, aes(x = parental_education_level, y = exam_score, fill = parental_education_level)) +
  geom_boxplot() +
  labs(title = "Puntuación del Examen según Nivel Educativo de los Padres",
       x = "Nivel Educativo de los Padres",
       y = "Puntuación del Examen (%)") +
  theme_minimal()
```




```
ggplot(datos, aes(x = part_time_job, y = exam_score, fill = part_time_job)) +
  geom_boxplot() +
  labs(title = "Puntuación del Examen según si Tienen Trabajo",
       x = "¿Tiene Trabajo de Medio Tiempo?",
       y = "Puntuación del Examen (%)") +
  theme_minimal()
```




```r
ggplot(datos, aes(x = parental_education_level, fill = internet_quality)) +
  geom_bar(position = "fill") + # "fill" muestra proporciones (porcentajes)
  labs(title = "Proporción de Calidad de Internet por Nivel Educativo Paterno",
       x = "Nivel Educativo de los Padres",
       y = "Proporción de Estudiantes",
       fill = "Calidad de Internet") +
  scale_y_continuous(labels = scales::percent) + # Eje Y en porcentaje
  theme_minimal()
```



Observaciones de los datos

El conjunto de datos está balanceado en cuanto a género, calidad de la dieta, calidad del internet y participación en actividades extracurriculares. La mayoría de los estudiantes no tiene un trabajo a tiempo parcial, y el nivel educativo más común de los padres es la preparatoria (High School). Las horas de estudio, el uso de redes sociales y los puntajes de los exámenes varían considerablemente entre los estudiantes.

Reducción de dimensión

Método PCA

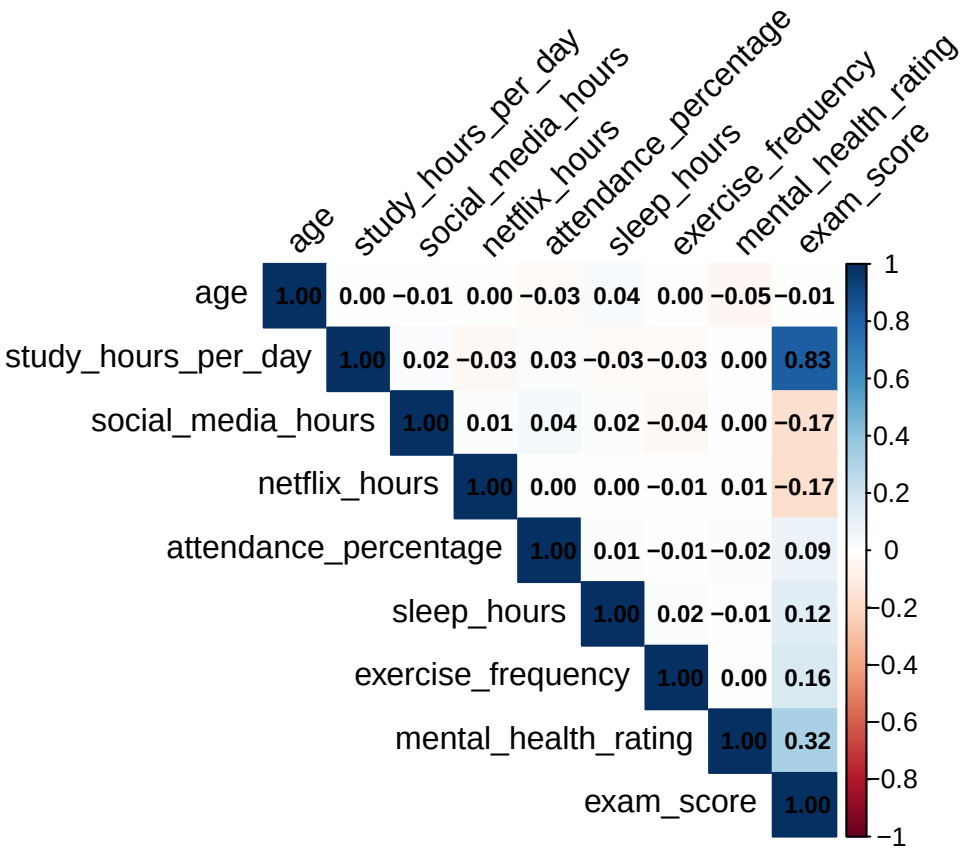
Antes de aplicar PCA, es crucial preparar los datos y entender los supuestos del algoritmo. El PCA solo puede aplicarse a variables numéricas y asume linealidad en la estructura de los datos.

Supuestos de PCA:

- Linealidad: PCA busca relaciones lineales entre las variables. Si las relaciones son no lineales, es posible que el PCA no capture la estructura subyacente de manera efectiva.
- Variabilidad: El método de PCA se basa en la varianza de los datos. Una mayor varianza en una dirección específica indica que esa dirección es importante y por lo tanto, un posible componente principal.
- Variables numéricas: PCA funciona con variables numéricas. Las variables categóricas deben ser transformadas a un formato numérico.

Para empezar, hagamos un pequeño análisis de la correlación de las variables numéricas.

```
numeric_vars <- habits %>%
  select(age, study_hours_per_day, social_media_hours, netflix_hours,
         attendance_percentage, sleep_hours, exercise_frequency,
         mental_health_rating, exam_score)
cor_matrix <- cor(numeric_vars)
corrplot(cor_matrix, method = "color", type = "upper",
         addCoef.col = "black", tl.col = "black", tl.srt = 45,  number.cex = 0.75)
```



Del gráfico, podemos notar que la correlación entre las caracterísitcas de los alumnos no está correlacionada. Por lo tanto, no esperamos tener un buen desempeño al aplicar el algoritmo de PCA.

Aplicando esto mismo para pasando las variables categóricas a datos numéricos usando variables dummy. Nótese que no usamos la función step_dummy para asegurarnos que la interpretación de cada variable se preserve, por ejemplo en la calidad de la dieta asociamos una dieta `Poor` con el número 1 y una dieta con `Good` al valor 3.

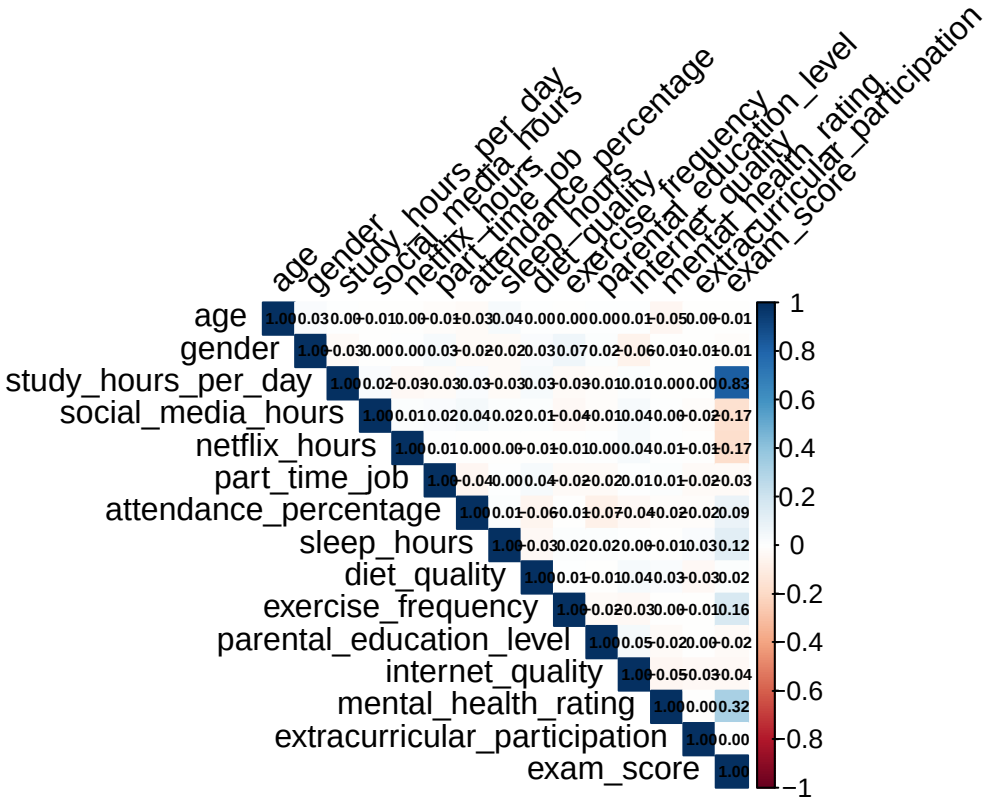
```
datos_convertidos <- habits |>
  mutate(
    # Variables binarias (0/1)
    gender = if_else(gender == "Male", 1, 0),
    part_time_job = if_else(part_time_job == "Yes", 1, 0),
    extracurricular_participation = if_else(extracurricular_participation == "Yes", 1, 0),
    diet_quality = case_when(
      diet_quality == "Poor"    ~ 1,
```

```
      diet_quality == "Fair"    ~ 2,
      diet_quality == "Good"    ~ 3,
      TRUE ~ NA_real_ # Por si hay otros valores
    ),

    internet_quality = case_when(
      internet_quality == "Poor"    ~ 1,
      internet_quality == "Average" ~ 2,
      internet_quality == "Good"    ~ 3,
      TRUE ~ NA_real_
    ),

    parental_education_level = case_when(
      parental_education_level == "None"        ~ 0,
      parental_education_level == "High School" ~ 1,
      parental_education_level == "Bachelor"    ~ 2,
      parental_education_level == "Master"      ~ 3,
      TRUE ~ NA_real_
    )
  )

cor_matrix <- cor(datos_convertidos)
corrplot(cor_matrix, method = "color", type = "upper",
         addCoef.col = "black", tl.col = "black", tl.srt = 45,  number.cex = 0.5)
```



Del gráfico anterior, tenemos el mismo resultado usando las variables categóricas como dummies. Por la

misma razón, no esperamos obtener un buen desempeño del algoritmo PCA en nuestros datos.

Aplicamos PCA en los datos estandarizados omitiendo la variable objetivo que es exam_score

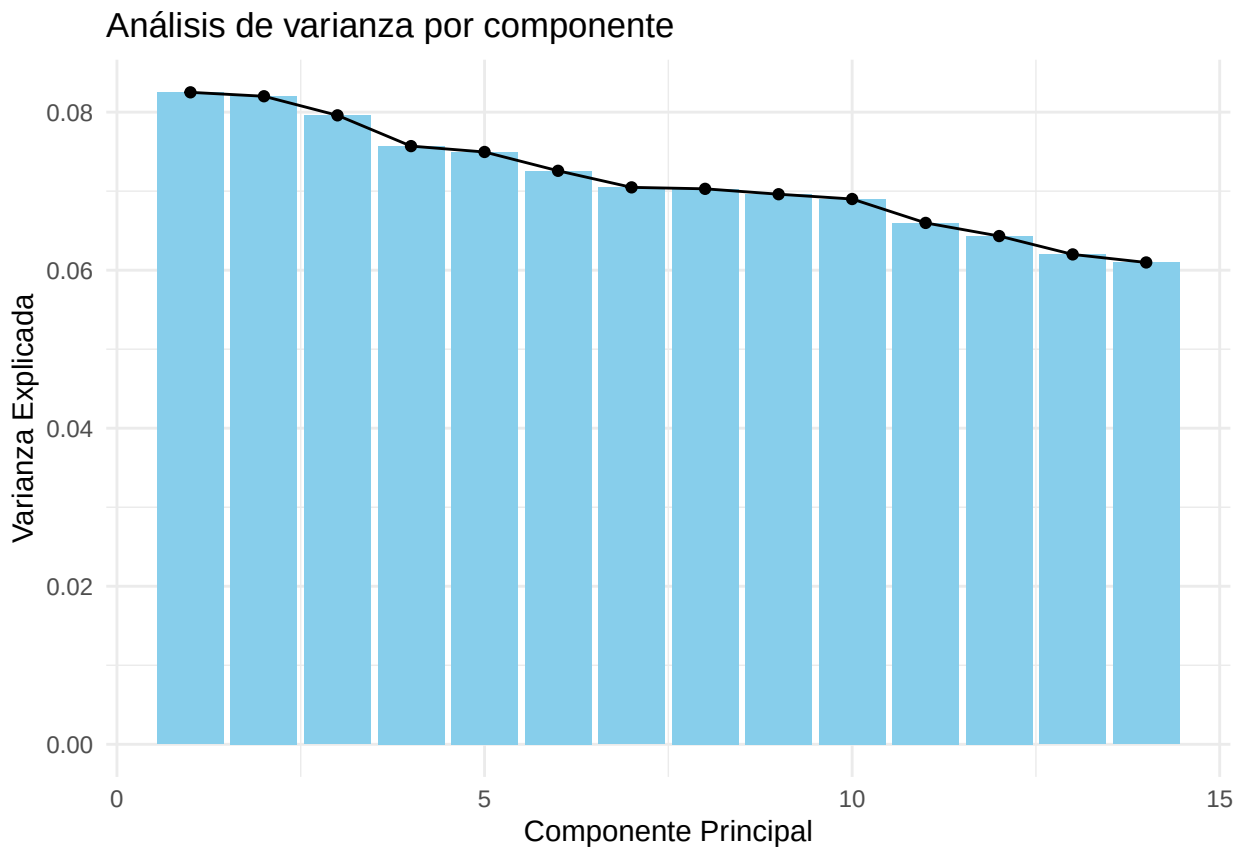
```
datos_pca <- datos_convertidos[, names(datos_convertidos) != "exam_score"]
pca_fit <- datos_pca |>
  scale() |>
  prcomp()

summary(pca_fit)
```

```
## Importance of components:
##                           PC1     PC2     PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.07474 1.07151 1.05561 1.0295 1.02440 1.00799 0.99333
## Proportion of Variance 0.08251 0.08201 0.07959 0.0757 0.07496 0.07257 0.07048
## Cumulative Proportion  0.08251 0.16451 0.24411 0.3198 0.39477 0.46734 0.53782
##                           PC8     PC9   PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.99203 0.98721 0.9829 0.96111 0.94892 0.93163 0.92393
## Proportion of Variance 0.07029 0.06961 0.0690 0.06598 0.06432 0.06199 0.06097
## Cumulative Proportion  0.60811 0.67773 0.7467 0.81271 0.87703 0.93903 1.00000
```

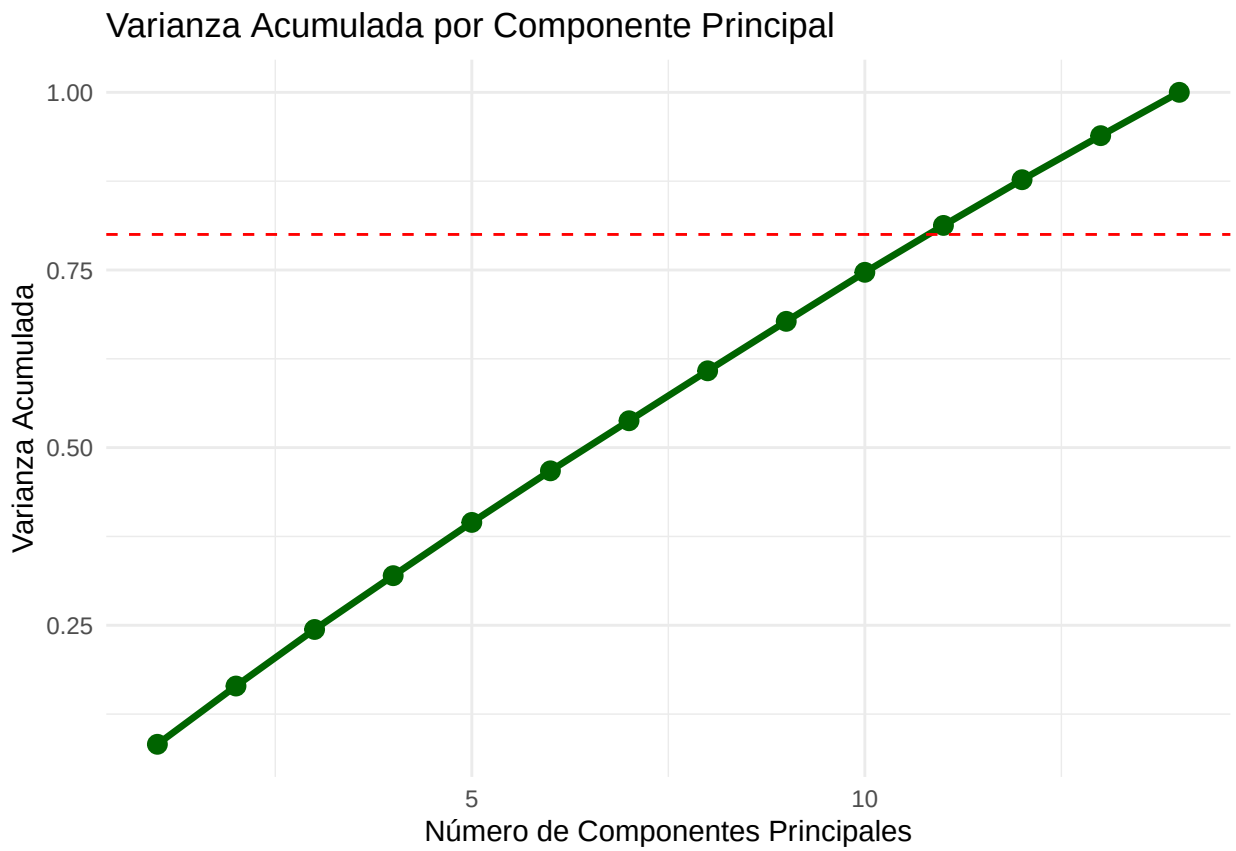
```
variance_explained <- pca_fit$sdev^2 / sum(pca_fit$sdev^2)
cumulative_variance <- cumsum(variance_explained)
```

```
data.frame(
  PC = 1:length(variance_explained),
  Variance = variance_explained
) %>%
  ggplot(aes(x = PC, y = Variance)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  geom_point() +
  geom_line() +
  labs(title = "Análisis de varianza por componente",
       x = "Componente Principal",
       y = "Varianza Explicada") +
  theme_minimal()
```




```
data.frame(
  PC = 1:length(cumulative_variance),
  Cumulative_Variance = cumulative_variance
) |>
  ggplot(aes(x = PC, y = Cumulative_Variance)) +
  geom_line(color = "darkgreen", size = 1.2) +
  geom_point(color = "darkgreen", size = 3) +
  labs(title = "Varianza Acumulada por Componente Principal",
       x = "Número de Componentes Principales",
       y = "Varianza Acumulada") +
  geom_hline(yintercept = 0.80, linetype = "dashed", color = "red") +
  theme_minimal()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



De la aplicación del algoritmo PCA, podemos notar que con aproximadamente 11 componentes principales, se puede explicar más del 80% de la varianza total, lo que significa que el espacio de 15 variables originales puede ser reducido significativamente a un subespacio de 11 dimensiones, manteniendo la mayor parte de la variabilidad de los datos. Lo cual indica que no es un buen desempeño del algoritmo PCA para nuetros datos.