Reporte de Analisis Detallado

Carlos Merino

07 de Febrero, 2026

Contents

- Extracción e Importación de Datos

La metodología de obtención y carga de datos se realizó en los siguientes pasos:

- Extracción de la Fuente

Primero se descargan los datos desde https://www.kaggle.com/ , escogiendo la base de datos "student_exam_scores.csv". Luego se guarda escogiendo cualquier directorio de alguna carpeta, en este caso, se escogio la ruta:

```
"C:/Users/carlo/Desktop/Proyectos propios/analisis con reporte de datos/Reporte de Datos/student_exam_s
```

- Importación a R

Despues se importa esta base de datos al entorno de R con el siguiente codigo:

```
data <- read_csv("student_exam_scores.csv")
```

Donde:

Elemento	Tipo	Descripción
`data`	Objeto	Es el nombre con el que guardaremos esta base de datos.
`read_csv()`	Función	Función que lee archivos en formato CSV, reconociendo los tipos de datos de las columnas.
`(student_exam...)`	Directorio	Ubicacion del archivo

- Limpieza basica y estandarizacion

Este análisis se basa en el conjunto de datos de puntajes estudiantiles, el cual se ha limpiado y estandarizado.

- Carga de Librerías

Para la estandarización de nombres de columnas, se utilizó la librería janitor.

- Proceso de Estandarización

Se aplicó la función clean_names() para estandarizar los nombres de todas las columnas, y la función mutate() para asegurar el formato de la variable clave del estudiante.

El código utilizado fue:

```
data <- data %>%
  janitor::clean_names() %>%
  mutate(student_id = as.character(student_id))
```

Donde:

Elemento	Tipo	Descripción
`data`	Objeto	Estandariza los nombres de todas las columnas (ej. de 'Previous Scores' a 'previous_scores').
`janitor::clean_names()`	Limpieza	Función que lee archivos en formato CSV, reconociendo los tipos de datos de las columnas.
`mutate()`	Transformación	Modifica la columna student_id para asegurar que se almacene como un carácter.

- Exploracion inicial y estadistica descriptiva

Esta etapa se realiza sobre el conjunto de datos transformado (data_clean) para obtener un primer vistazo a la estructura, validar el tipo de dato de las nuevas variables y resumir las características centrales de las variables numéricas.

- glimpse(data_clean)

La función glimpse() proporciona un resumen conciso de las filas, columnas y, lo más importante, el tipo de dato de cada variable (dbl, chr). Esto confirma que las transformaciones (como la de student_id a carácter) se realizaron correctamente.

```
glimpse(data)
```

```
## Rows: 200
## Columns: 6
## $ student_id         <chr> "S001", "S002", "S003", "S004", "S005", "S006", "S0~
## $ hours_studied      <dbl> 8.0, 1.3, 4.0, 3.5, 9.1, 8.4, 10.8, 2.0, 5.6, 1.3, ~
## $ sleep_hours        <dbl> 8.8, 8.6, 8.2, 4.8, 6.4, 5.1, 6.0, 4.3, 5.9, 8.9, 5~
## $ attendance_percent <dbl> 72.1, 60.7, 73.7, 95.1, 89.8, 58.5, 54.2, 75.8, 81.~
## $ previous_scores    <dbl> 45, 55, 86, 66, 71, 75, 88, 55, 84, 70, 81, 85, 71,~
## $ exam_score         <dbl> 30.2, 25.0, 35.8, 34.0, 40.3, 35.7, 37.9, 18.3, 34.~
```

- summary(data_clean)

La función summary() es crucial, ya que calcula métricas descriptivas esenciales (media, mediana, cuartiles) para las variables continuas.

```
summary(data)
```

```
##   student_id        hours_studied     sleep_hours    attendance_percent
##  Length:200         Min.   : 1.000   Min.   :4.000   Min.   : 50.30
##  Class :character   1st Qu.: 3.500   1st Qu.:5.300   1st Qu.: 62.20
##  Mode  :character   Median : 6.150   Median :6.700   Median : 75.25
##                     Mean   : 6.325   Mean   :6.622   Mean   : 74.83
##                     3rd Qu.: 9.000   3rd Qu.:8.025   3rd Qu.: 87.42
##                     Max.   :12.000   Max.   :9.000   Max.   :100.00
```

```
##  previous_scores   exam_score
##  Min.   :40.0    Min.   :17.10
##  1st Qu.:54.0    1st Qu.:29.50
##  Median :67.5    Median :34.05
##  Mean   :66.8    Mean   :33.95
##  3rd Qu.:80.0    3rd Qu.:38.75
##  Max.   :95.0    Max.   :51.30
```

- head(data_clean)

La función head() permite visualizar las primeras filas del dataset. Esto sirve como una validación rápida para asegurar que las nuevas variables categóricas (pass_fail, previos_cat) se han creado con los valores esperados.

```
head(data)
```

```
## # A tibble: 6 x 6
##   student_id hours_studied sleep_hours attendance_percent previous_scores
##   <chr>              <dbl>       <dbl>              <dbl>           <dbl>
## 1 S001                 8           8.8               72.1              45
## 2 S002                 1.3         8.6               60.7              55
## 3 S003                 4           8.2               73.7              86
## 4 S004                 3.5         4.8               95.1              66
## 5 S005                 9.1         6.4               89.8              71
## 6 S006                 8.4         5.1               58.5              75
## # i 1 more variable: exam_score <dbl>
```

- Transformacion de los datos a traves de variables nuevas

A partir del dataset limpio, se crean seis nuevas variables categóricas y de ratio (como el nivel de aprobación, el nivel previo y el balance estudio/sueño) esenciales para el análisis de rendimiento.

```
data_clean <- data %>%
  mutate(
    study_per_sleep = hours_studied / sleep_hours,
    previos_cat = case_when(
      previous_scores >= 80 ~ "Alto",
      previous_scores >= 60 ~ "Medio",
      TRUE ~ "Bajo"
    ),
    exam_level = case_when(
      exam_score >= 40 ~ "Alto",
      exam_score >= 30 ~ "Medio",
      TRUE ~ "Bajo"
    ),
    pass_fail = if_else(exam_score >= 30, "Aprobado", "Reprobado"),
    exam_zscore = (exam_score - mean(exam_score, na.rm = TRUE)) / sd(exam_score, na.rm = TRUE),
    exam_percentile = percent_rank(exam_score) * 100
  )
```

Donde:

Elemento	Tipo	Descripción
`student_id`	Carácter	Asegura que el ID sea tratado como texto, no como valor numérico.
`study_per_sleep`	Ratio (Numérico)	Mide el balance entre el tiempo dedicado al estudio y el descanso (clave para el gráfico de burbujas).
`previos_cat`	Categórica	Clasifica el historial académico en Alto, Medio o Bajo para análisis comparativos.
`exam_level`	Categórica	Clasifica el puntaje del examen en Alto, Medio o Bajo.
`pass_fail`	Binaria	Determina el estado de aprobación/reprobación (Aprobado >= 30, Reprobado < 30). Crucial para la correlación.
`exam_zscore`	Numérico	Estandariza el puntaje para medir qué tan lejos está de la media general.
`exam_percentile`	Numérico	Mide el porcentaje de estudiantes que obtuvieron un puntaje igual o inferior.

- Graficos de ggplot2

- Histograma

El **histograma** a continuación muestra la distribución de la frecuencia de los puntajes obtenidos, con una línea vertical indicando la **media general** del examen. Esto permite identificar rápidamente la concentración de resultados.

- Densidad por categoría de aprobación

El **gráfico de densidad** compara las distribuciones de puntajes para los estudiantes que **Aprobaron** y **Reprobaron**. La densidad es clave para visualizar la separación entre ambos grupos respecto al umbral de aprobación.

- Boxplot

Este **gráfico de cajas y bigotes (Boxplot)** evalúa la distribución del puntaje del examen actual segmentado por la **categoría de puntaje previo** (Bajo, Medio, Alto). La caja muestra la mediana y los cuartiles de cada grupo.

- Gráfico de dispersión

El **gráfico de dispersión** a continuación muestra la fuerte **correlación positiva** entre la cantidad de horas dedicadas al estudio y el puntaje obtenido en el examen, diferenciando la tendencia por el resultado final (Aprobado/Reprobado).

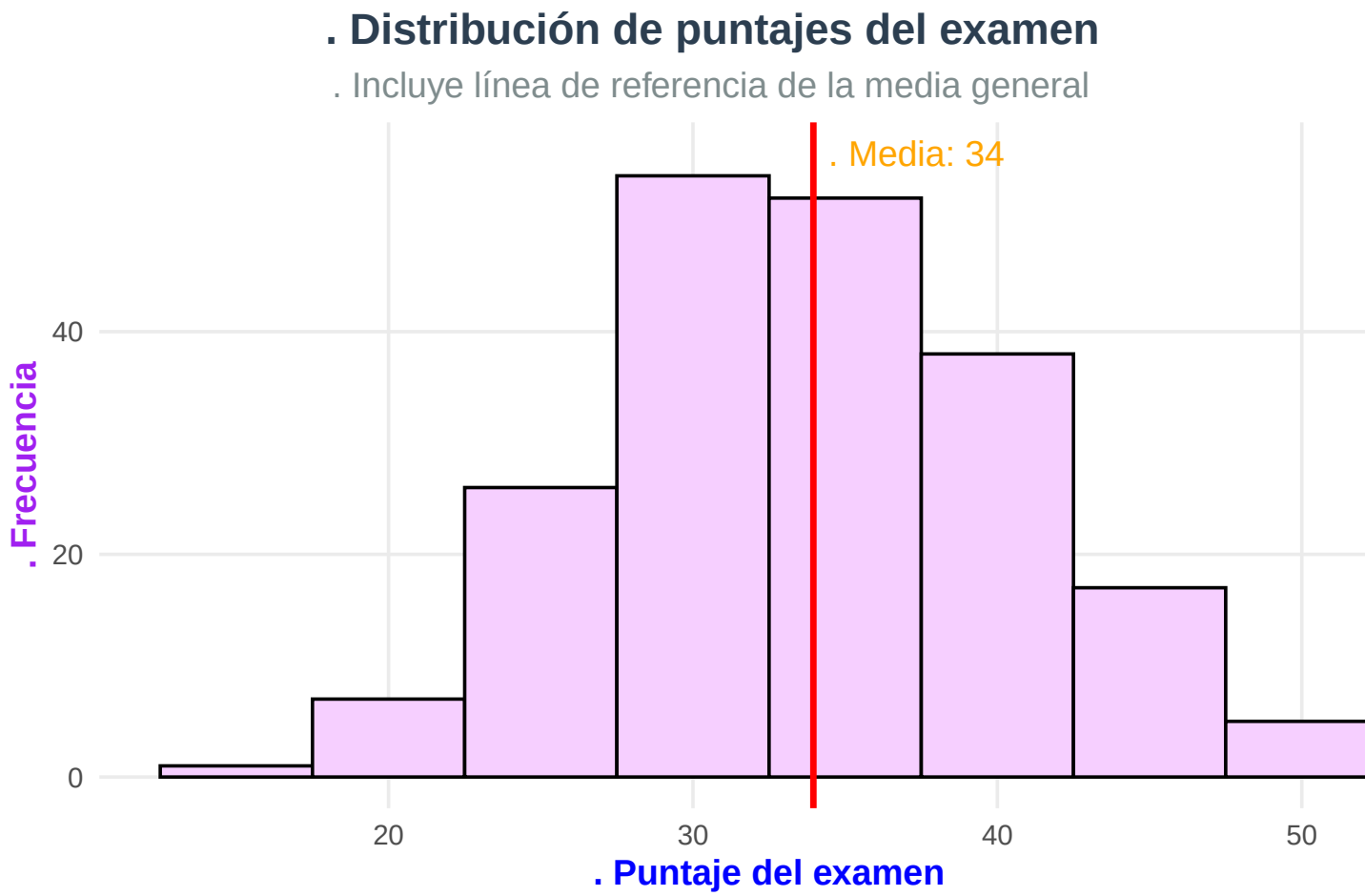

Figure 1: Gráfico de Histograma: Relación entre Frecuencias y Puntaje

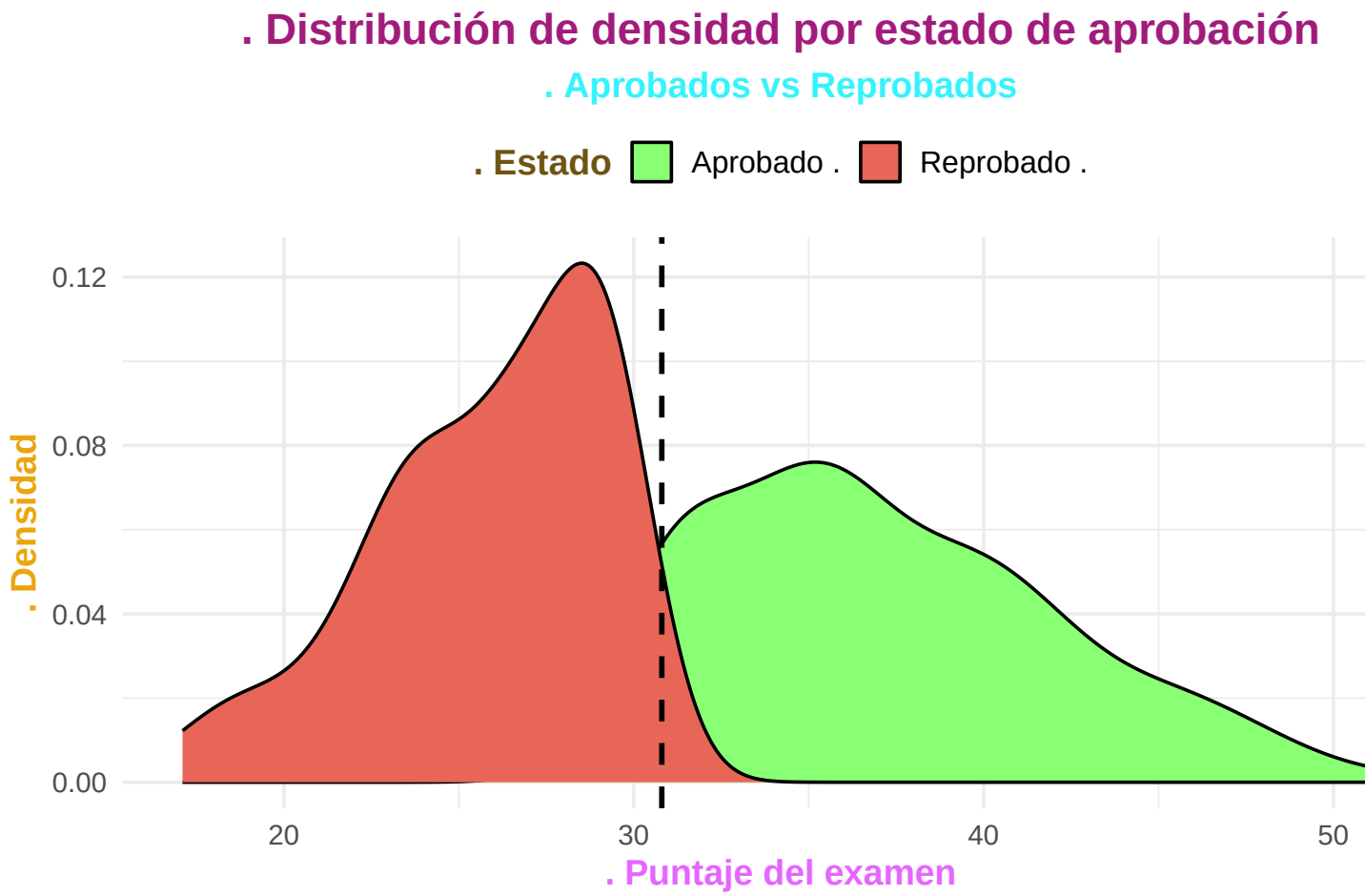

Figure 2: Gráfico de Densidad: Relación entre estado de aprobación

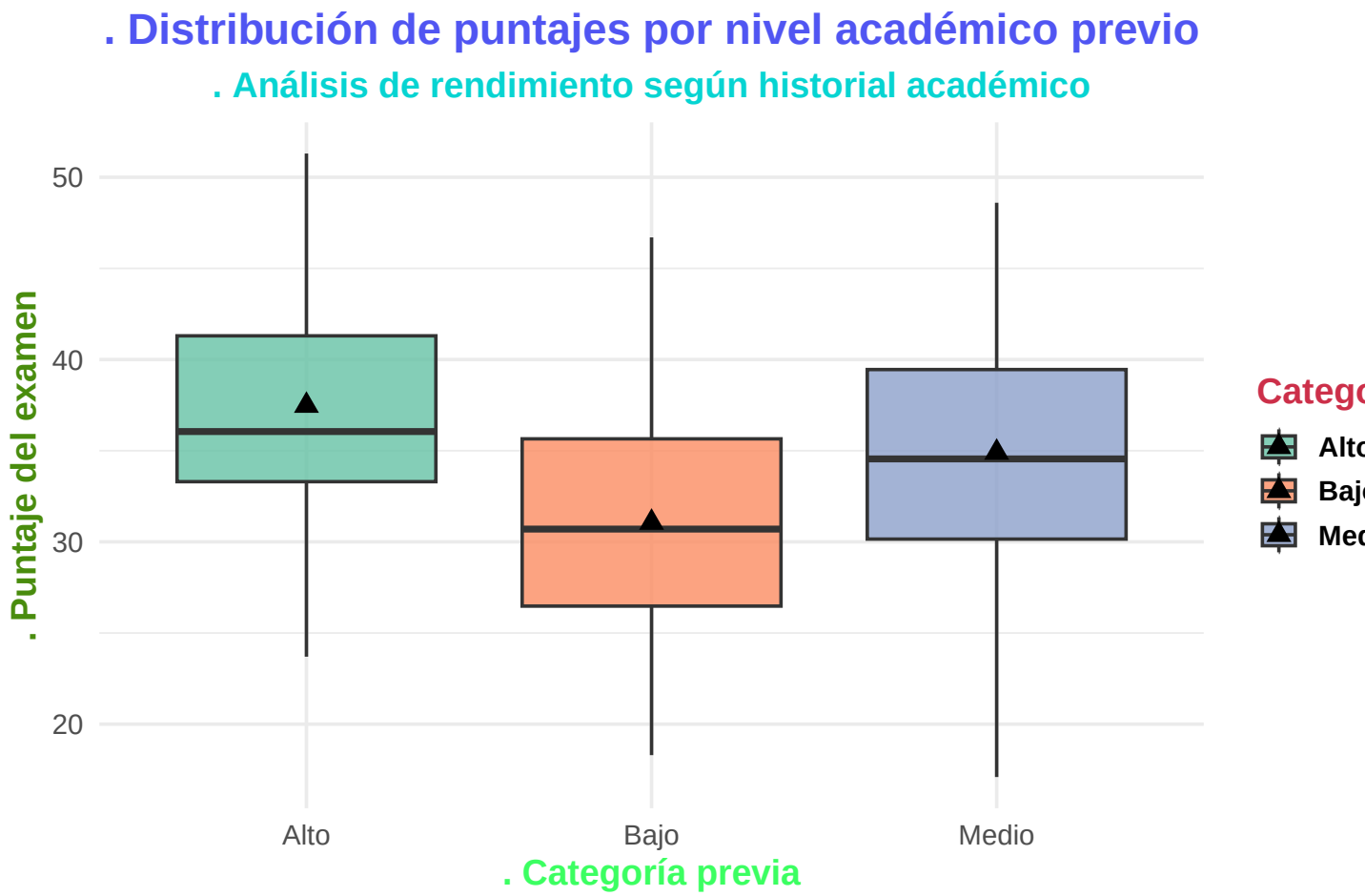

Figure 3: Gráfico de Boxplot: Distribución de puntajes

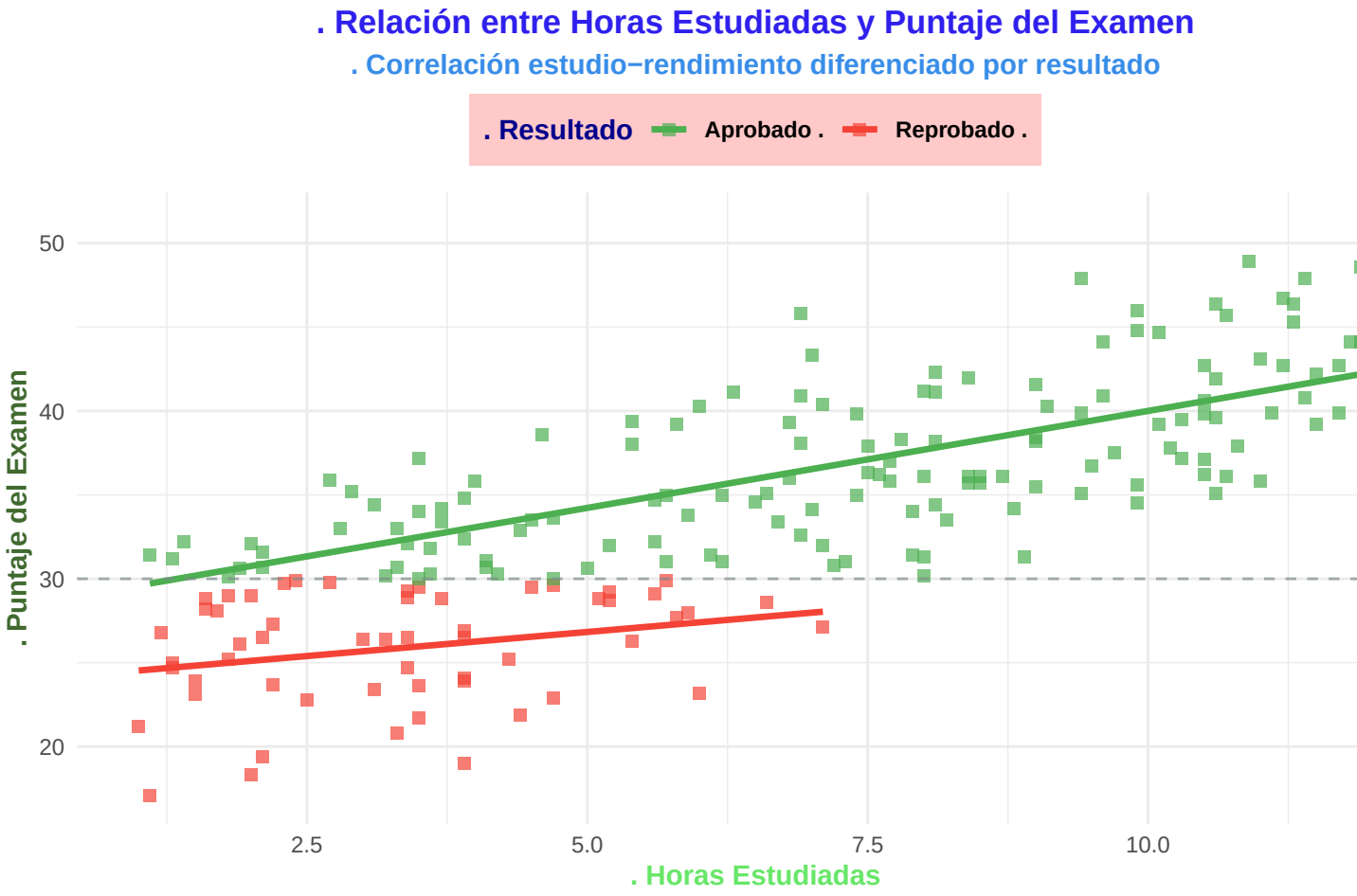

Figure 4: Gráfico de Dispersión: Relación entre horas estudiadas y puntaje del examen

- **Gráfico de barras**

El **gráfico de barras** muestra la **frecuencia absoluta y relativa** de estudiantes clasificados en los niveles de desempeño (Bajo, Medio, Alto) creados a partir del puntaje final del examen.

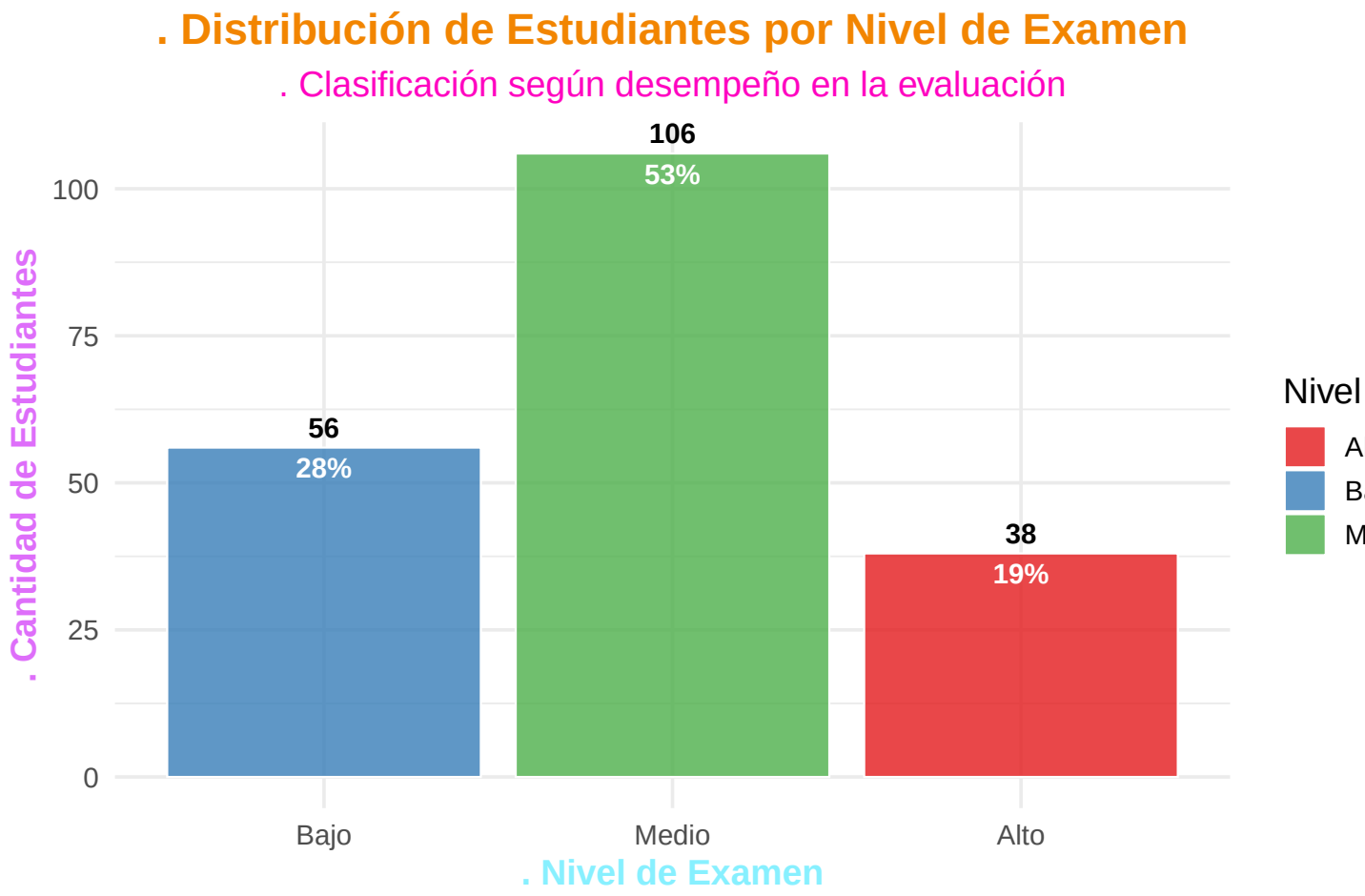

Figure 5: Gráfico de Barras: Distribución de Estudiantes por Nivel de Examen

- **Análisis del ratio estudio/sueño**

El **gráfico de burbujas** evalúa el **balance entre horas estudiadas y horas de sueño (Ratio)** en relación con el puntaje. El **tamaño de las burbujas** representa las horas de sueño, indicando si el descanso es un factor clave en el rendimiento final.

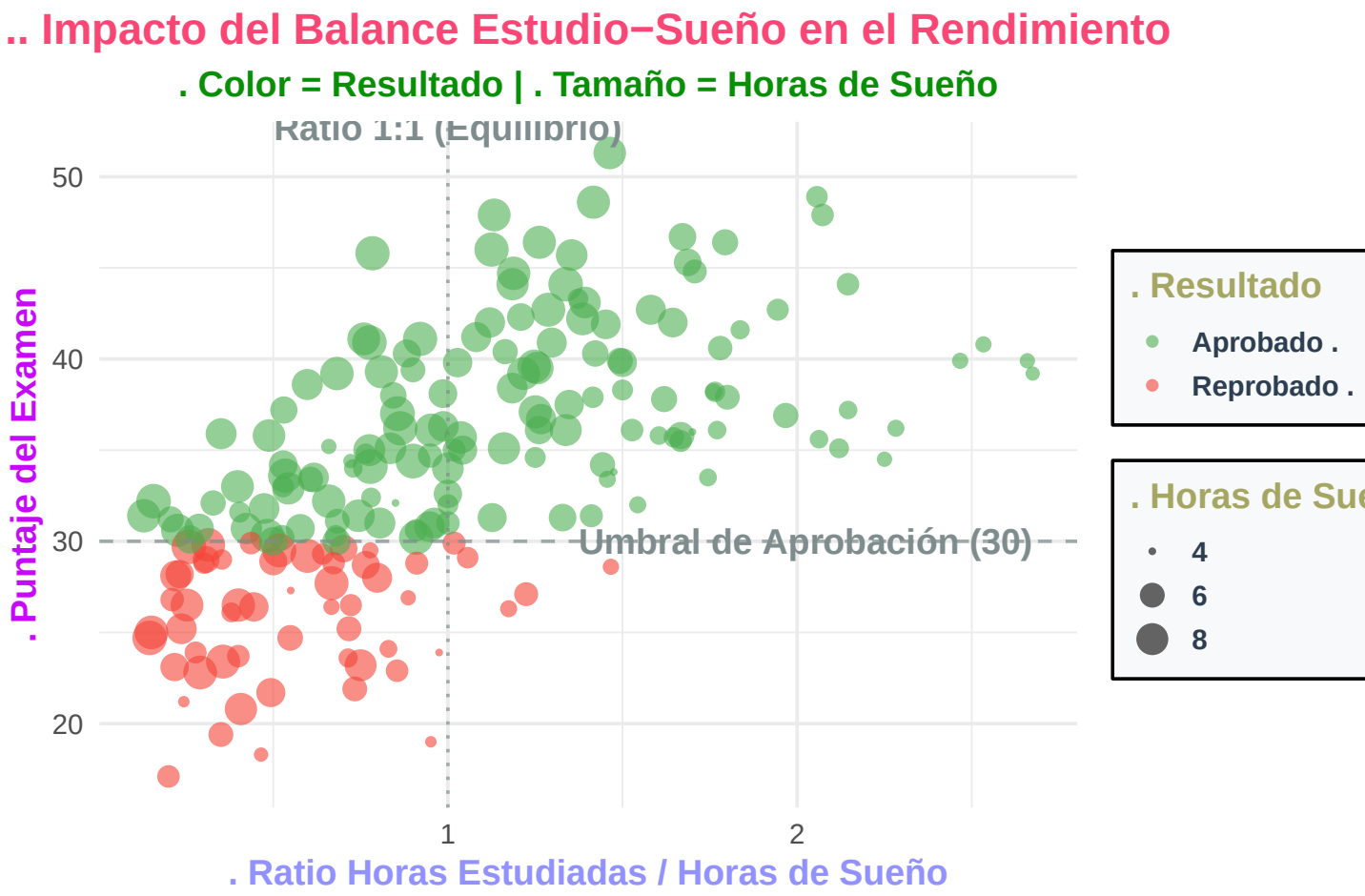

Figure 6: Gráfico de Análisis: ratio estudio/sueño