PR2 Tipologia y ciclo de vida de los datos

Alejandro González Barberá y Ferran Valverde Parera

2023-01-09

Contents

Descripción del dataset

Este conjunto de datos ha sido tomado de https://www.kaggle.com/datasets/whenamancodes/students-performance-in-exams, y está formado por las notas de los estudiantes en diferentes asignaturas junto a algunas características de cada estudiante. Es interesante analizar este conjunto de datos para poder analizar y sacar conclusiones sobre que aspectos pueden influir más en el rendimiento de los estudiantes. A partir de estas conclusiones se pueden llegar a predecir las notas y reconducir a tiempo en caso de esperar un mal resultado académico.

```
data <- read.csv('../data/exams.csv')
```

Integración y selección

En este caso no eliminaremos ningún atributo sino que añadiremos variables categóricas para las notas, es decir, suspenso, aprobado, notable, excelente. También añadiremos una columna para las medias de las tres asignaturas para cada estudiante.

```
data$math.result <- cut(data$math.score,
              breaks=c(0, 50, 70, 90, 100),
              labels=c('Fail', 'Sufficient', 'Good', 'Excellent'))
data$reading.result <- cut(data$reading.score,
              breaks=c(0, 50, 70, 90, 100),
              labels=c('Fail', 'Sufficient', 'Good', 'Excellent'))
data$writing.result <- cut(data$writing.score,
              breaks=c(0, 50, 70, 90, 100),
              labels=c('Fail', 'Sufficient', 'Good', 'Excellent'))
data$meangrade <- rowMeans(data[6:8])

head(data)
```

```
##   gender race.ethnicity parental.level.of.education        lunch
## 1   male        group A                 high school     standard
## 2 female        group D            some high school free/reduced
## 3   male        group E                some college free/reduced
## 4   male        group B                 high school     standard
## 5   male        group E           associate's degree     standard
## 6 female        group D                 high school     standard
##   test.preparation.course math.score reading.score writing.score math.result
## 1               completed         67            67            63  Sufficient
## 2                    none         40            59            55        Fail
## 3                    none         59            60            50  Sufficient
## 4                    none         77            78            68        Good
## 5               completed         78            73            68        Good
## 6                    none         63            77            76  Sufficient
##   reading.result writing.result meangrade
## 1     Sufficient     Sufficient  65.66667
## 2     Sufficient     Sufficient  51.33333
## 3     Sufficient           Fail  56.33333
## 4           Good     Sufficient  74.33333
## 5           Good     Sufficient  73.00000
## 6           Good           Good  72.00000
```

Limpieza de los datos

Antes de nada empezaremos corrigiendo los tipos de cada atributo en caso de que sea necesario.

```
sapply(data, class)
```

```
##                     gender              race.ethnicity
##                "character"                 "character"
## parental.level.of.education                       lunch
##                "character"                 "character"
##     test.preparation.course                  math.score
##                "character"                   "integer"
##               reading.score                writing.score
##                   "integer"                   "integer"
##                 math.result              reading.result
```

```
##                     "factor"                     "factor"
##              writing.result                   meangrade
##                     "factor"                    "numeric"
```

```
data$gender <- as.factor(data$gender)
data$race.ethnicity <- as.factor(data$race.ethnicity)
data$parental.level.of.education <- as.factor(data$parental.level.of.education)
data$lunch <- as.factor(data$lunch)
data$test.preparation.course <- as.factor(data$test.preparation.course)
```

Comprobamos que los cambios se han aplicado correctamente

```
summary(data)
```

```
##     gender    race.ethnicity     parental.level.of.education          lunch
##  female:483   group A: 79    associate's degree:203          free/reduced:348
##  male  :517   group B:205    bachelor's degree :112          standard    :652
##               group C:323    high school       :202
##               group D:262    master's degree   : 70
##               group E:131    some college      :222
##                              some high school  :191
##  test.preparation.course   math.score    reading.score writing.score
##  completed:335           Min.   : 13.0   Min.   : 27   Min.   : 23.00
##  none     :665           1st Qu.: 56.0   1st Qu.: 60   1st Qu.: 58.00
##                          Median : 66.5   Median : 70   Median : 68.00
##                          Mean   : 66.4   Mean   : 69   Mean   : 67.74
##                          3rd Qu.: 77.0   3rd Qu.: 79   3rd Qu.: 79.00
##                          Max.   :100.0   Max.   :100   Max.   :100.00
##      math.result     reading.result    writing.result   meangrade
##  Fail      :155   Fail      :113    Fail      :151    Min.   : 21.67
##  Sufficient:436   Sufficient:404    Sufficient:410    1st Qu.: 58.58
##  Good      :354   Good      :415    Good      :371    Median : 67.33
##  Excellent : 55   Excellent : 68    Excellent : 68    Mean   : 67.71
##                                                       3rd Qu.: 78.33
##                                                       Max.   :100.00
```

Con el resultado de arriba podemos comprobar también si existen valores en blanco o NA, pero con las siguientes funciones es más facil de apreciar.

```
#Valores NA
colSums(is.na(data))
```

```
##                       gender             race.ethnicity
##                            0                          0
## parental.level.of.education                      lunch
##                            0                          0
##     test.preparation.course                 math.score
##                            0                          0
##               reading.score                writing.score
##                            0                          0
##                 math.result             reading.result
```

```
##                            0                            0
##              writing.result                    meangrade
##                            0                            0
```

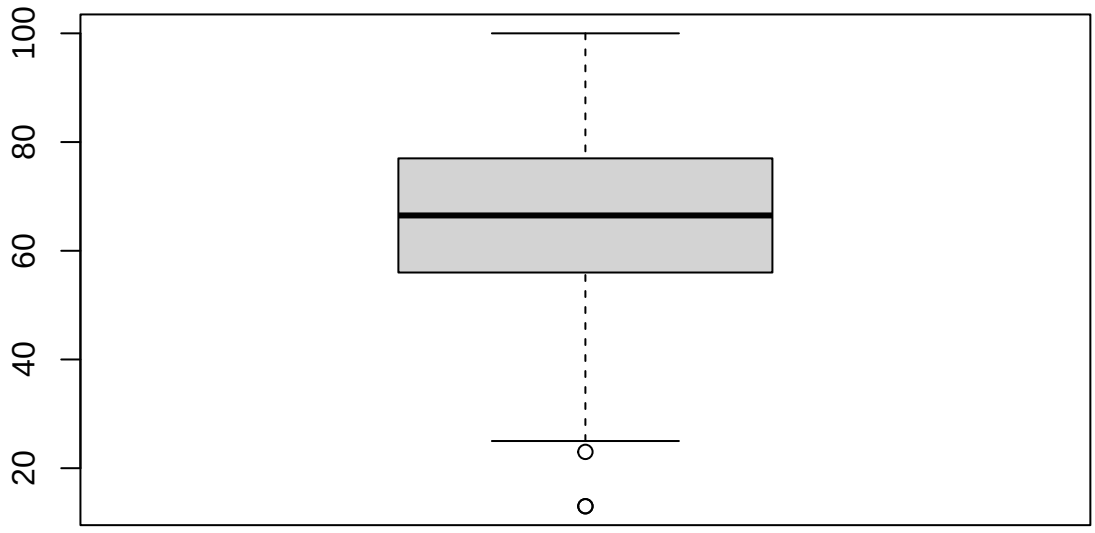
```
#Valores en blanco
colSums(data=="")
```

```
##                       gender              race.ethnicity
##                            0                            0
## parental.level.of.education                        lunch
##                            0                            0
##     test.preparation.course                   math.score
##                            0                            0
##                reading.score                writing.score
##                            0                            0
##                  math.result              reading.result
##                            0                            0
##              writing.result                    meangrade
##                            0                            0
```

Como podemos observar no encontramos ningún valor en blanco o NA.

Veamos si encontramos outliers en las variables numéricas.

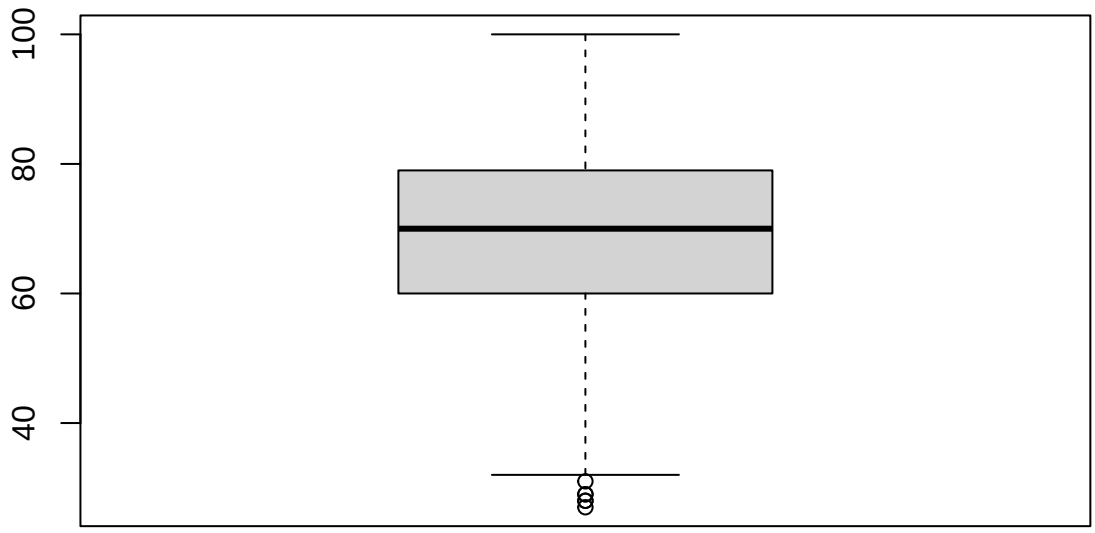
```
boxplot(data$math.score)
```



```
boxplot.stats(data$math.score)$out
```

```
## [1] 23 13 13
```

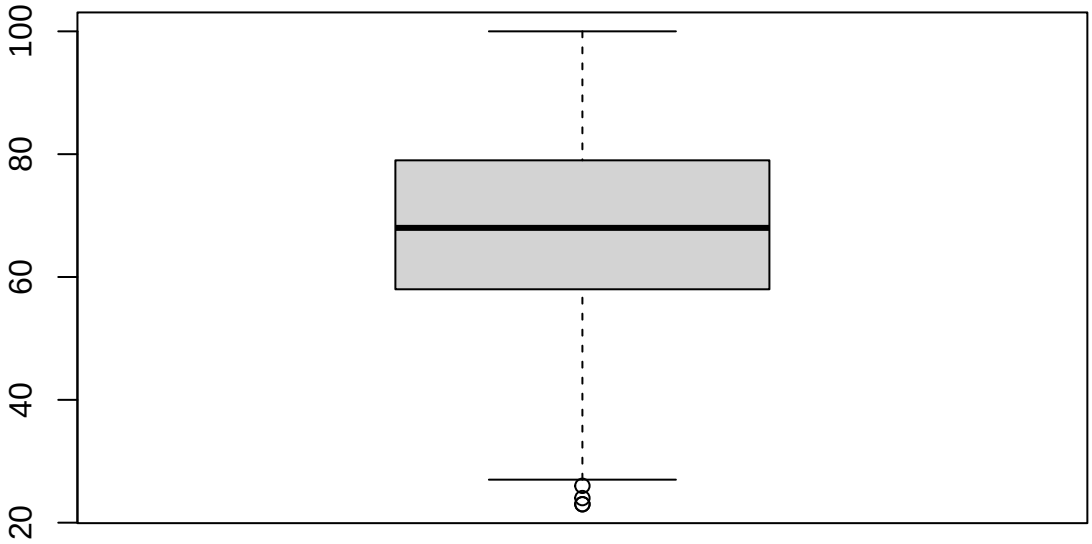
```
boxplot(data$reading.score)
```




```
boxplot.stats(data$reading.score)$out
```

```
## [1] 28 29 27 28 31 29
```

```
boxplot(data$writing.score)
```




```
boxplot.stats(data$writing.score)$out
```

```
## [1] 24 23 26 23
```

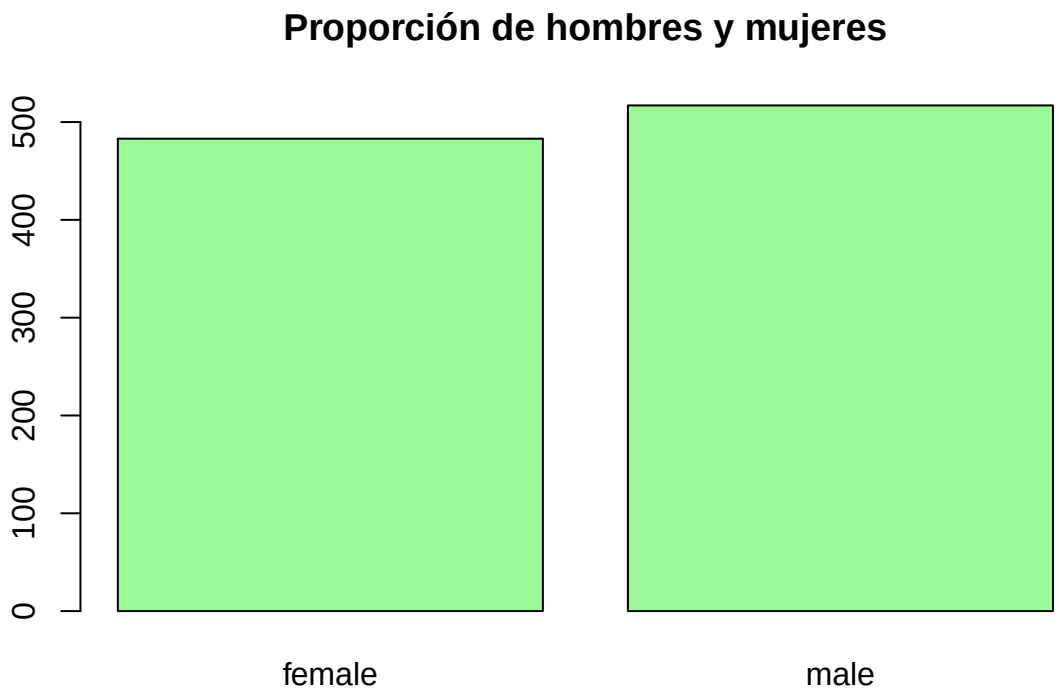
Vemos que si que hay unos pocos outliers para las notas de cada una de las asignaturas, podría ser conveniente eliminarlas pero al tratarse de una franja de notas cerrada (0-10) queremos analizar la totalidad de resultados, además no tenemos un motivo significante para eliminarlas.

Análisis y representación de los datos

Para analizar los datos, queremos comprobar mediante gráficos algunas de las relaciones y distribuciones de variables.

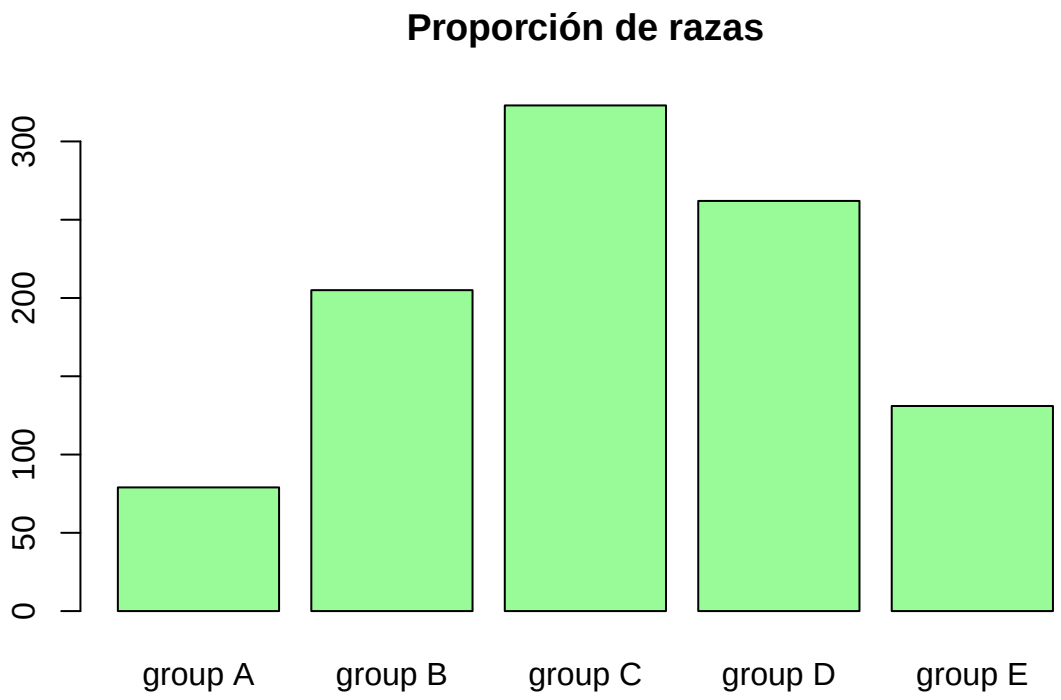
Distribuciones de los atributos

```
plot(data$gender, col = 'palegreen', main = 'Proporción de hombres y mujeres')
```



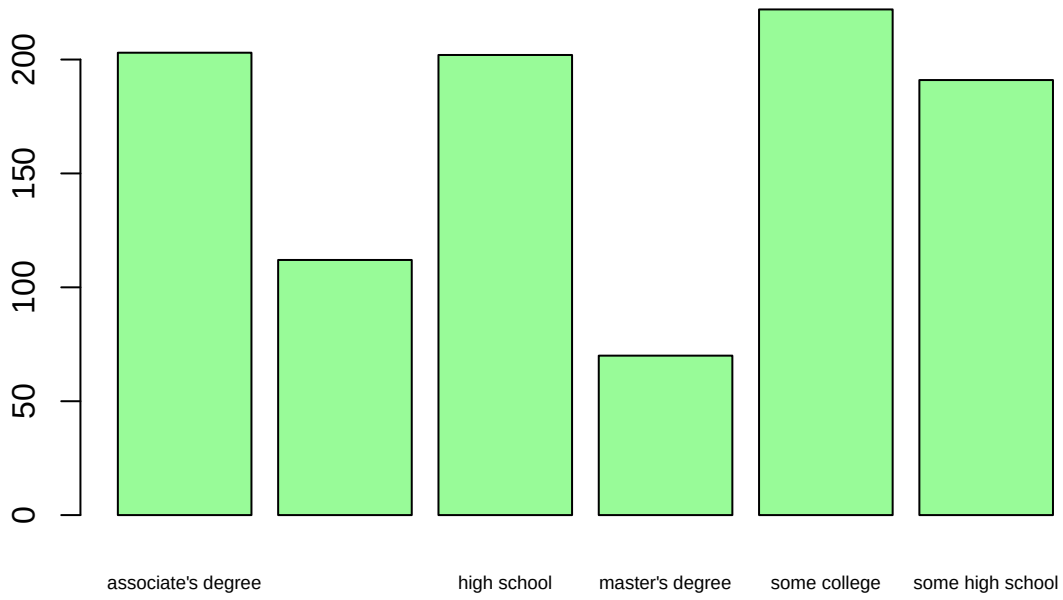
Hay aproximadamente el mismo número de hombres que de mujeres, aun que el total de hombres es ligeramente superior.

```
plot(data$race.ethnicity, col = 'palegreen', main = 'Proporción de razas')
```



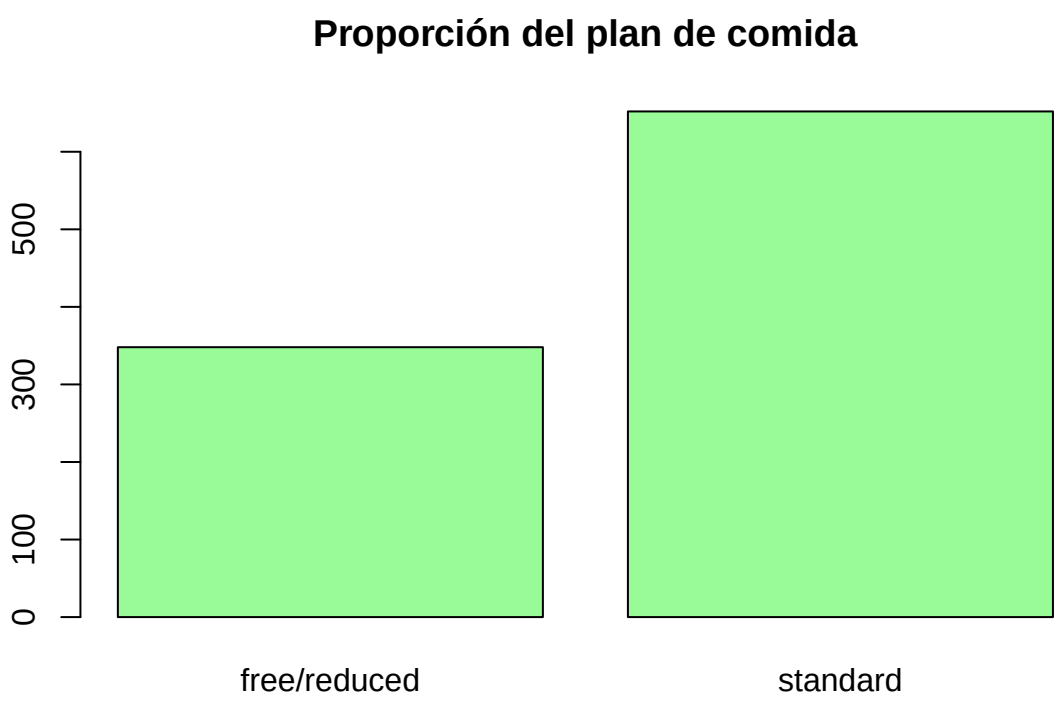
Como podemos apreciar, en este conjunto de datos se han especificado las razas de manera anónima. Esto puede ser interesante de cara a hacer los análisis sin suposiciones previas, aunque sería interesante conocer a que razas pertenece cada grupo una vez terminado el análisis.

```
plot(data$parental.level.of.education, cex.names=0.6, col = 'palegreen', main =
↪  'Proporción de estudios de los padres')
```



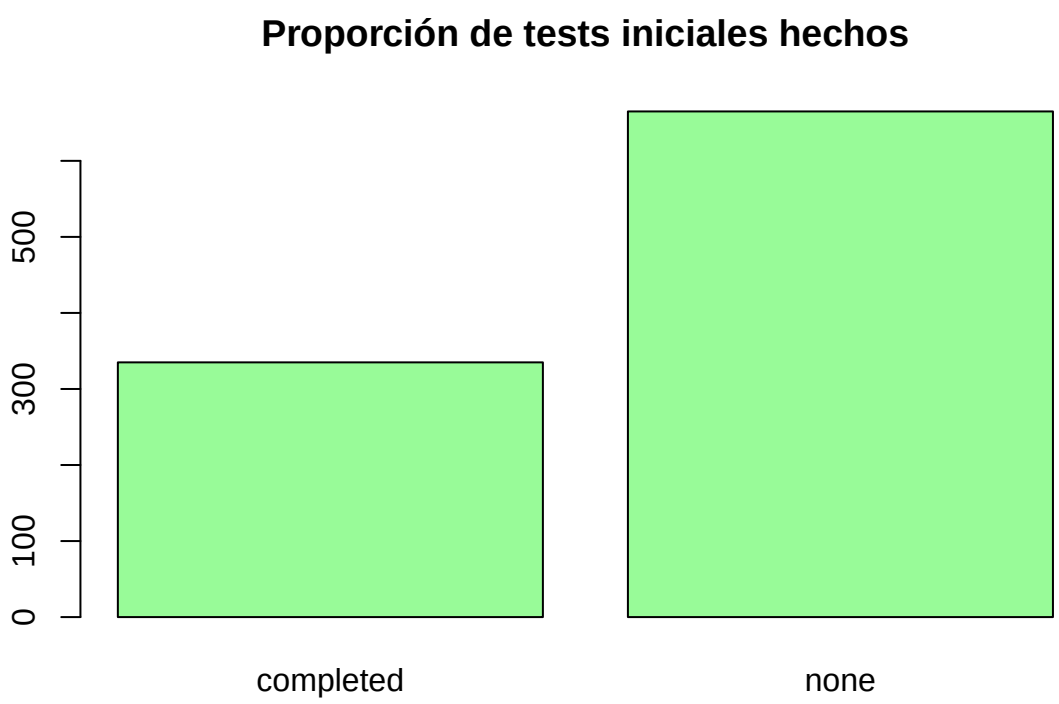
Podemos ver que muy pocos han llegado a hacer un máster como se podría llegar a esperar, pero también muy pocos han llegado a estudiar un grado.

```
plot(data$lunch, col = 'palegreen', main = 'Proporción del plan de comida')
```



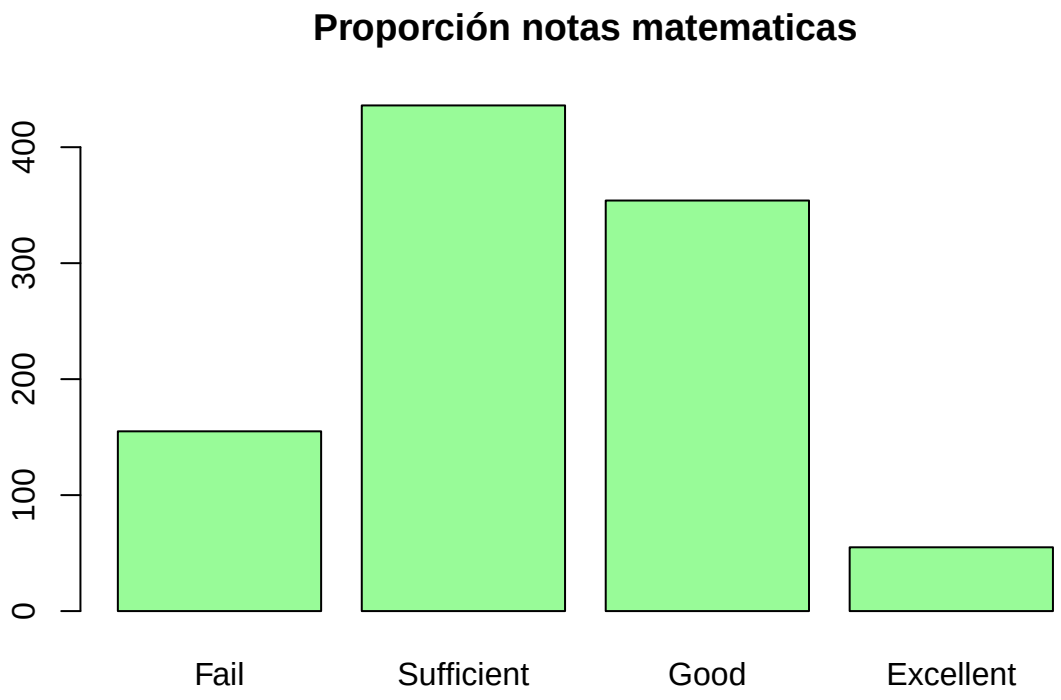
La mayoría tiene un plan estándar, pero muchos también tienen un plan reducido.

```
plot(data$test.preparation.course, col = 'palegreen', main = 'Proporción de tests
↪  iniciales hechos')
```

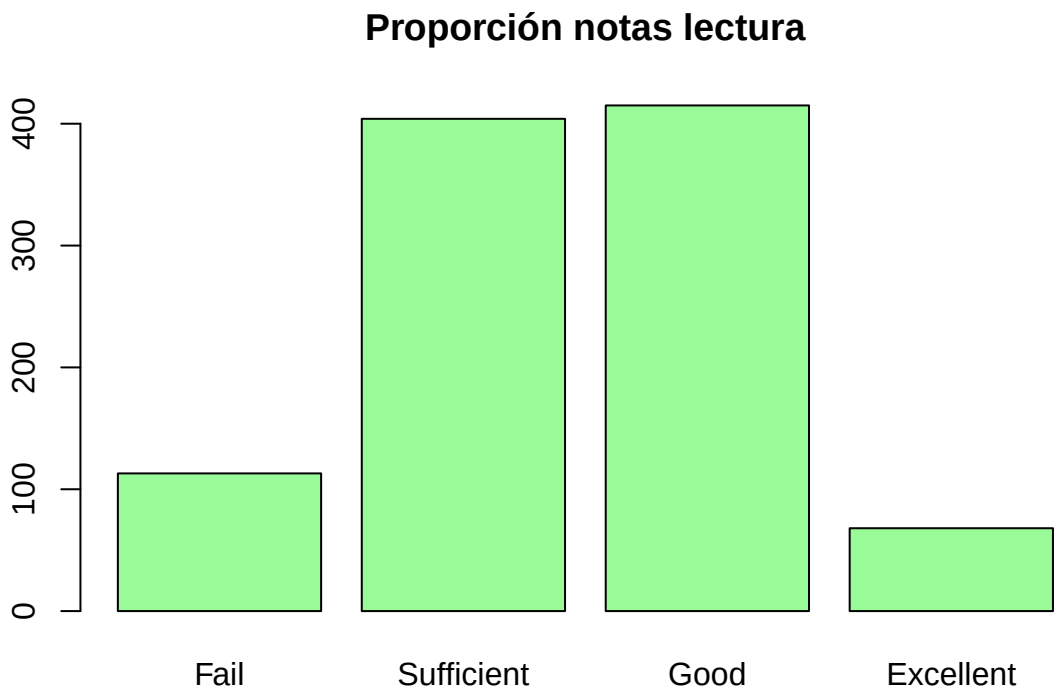


En este caso la mayoría no ha hecho el test de preparación.

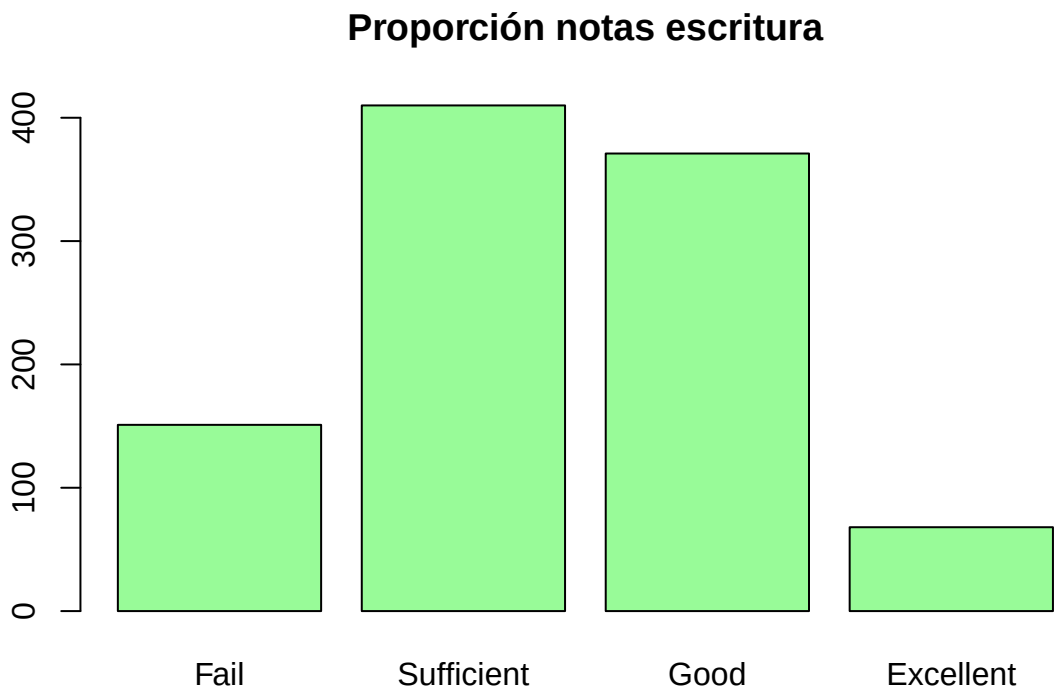
```
plot(data$math.result, col = 'palegreen', main = 'Proporción notas matematicas')
```




```
plot(data$reading.result, col = 'palegreen', main = 'Proporción notas lectura')
```



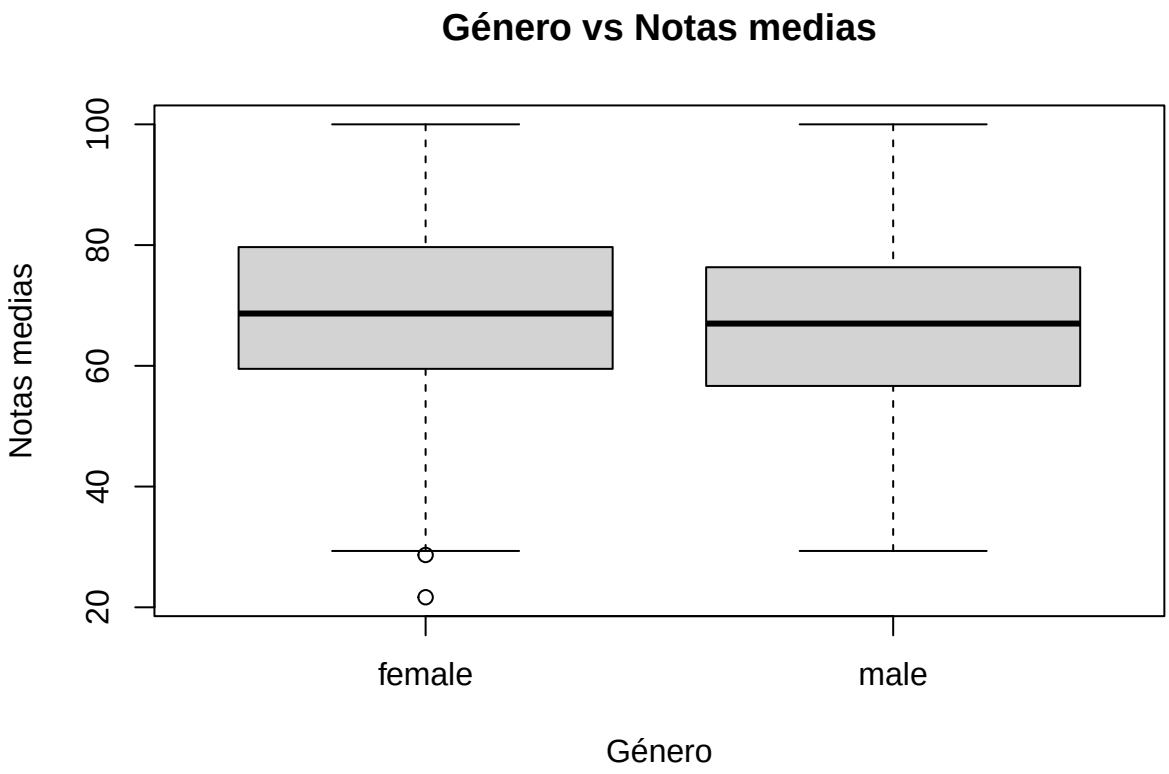

```
plot(data$writing.result, col = 'palegreen', main = 'Proporción notas escritura')
```



En las tres habilidades muestreadas vemos que la mayoría de notas se concentran en la zona central, es decir, están aprobados pero no con la máxima nota.

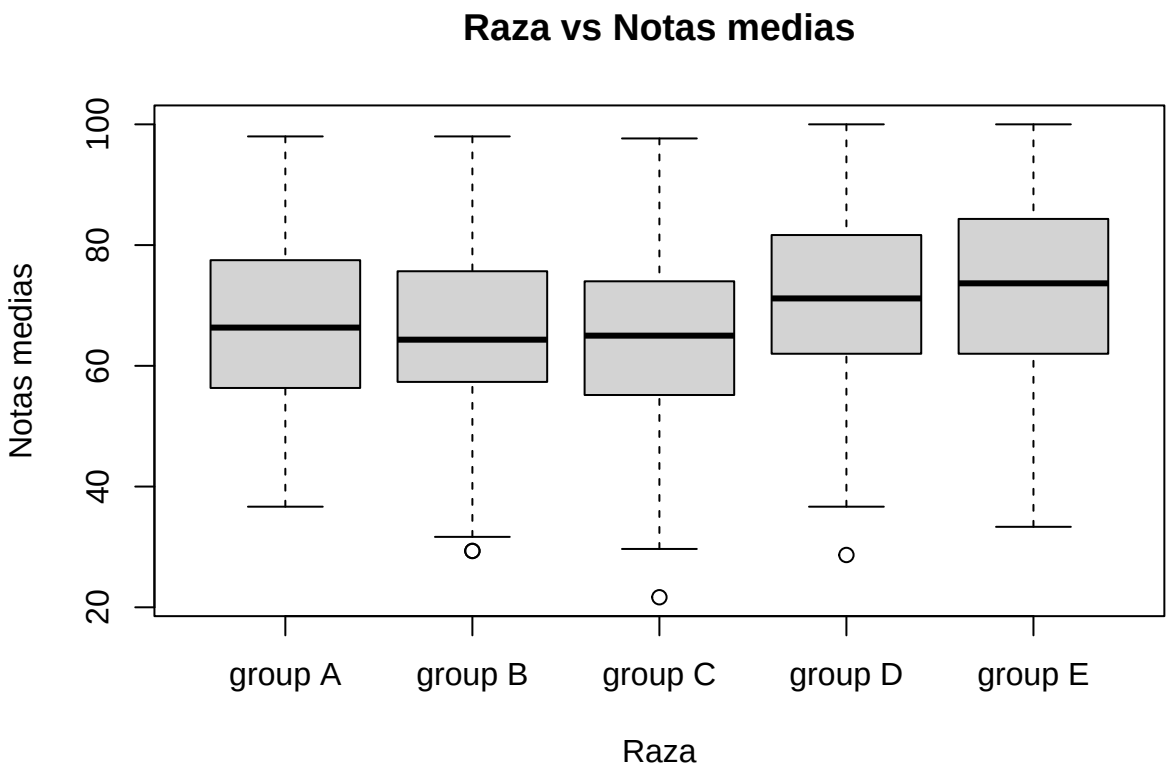
Relaciones entre atributos

```
plot(data$gender,data$meangrade, main = 'Género vs Notas medias', ylab = 'Notas medias',
↪  xlab = 'Género')
```



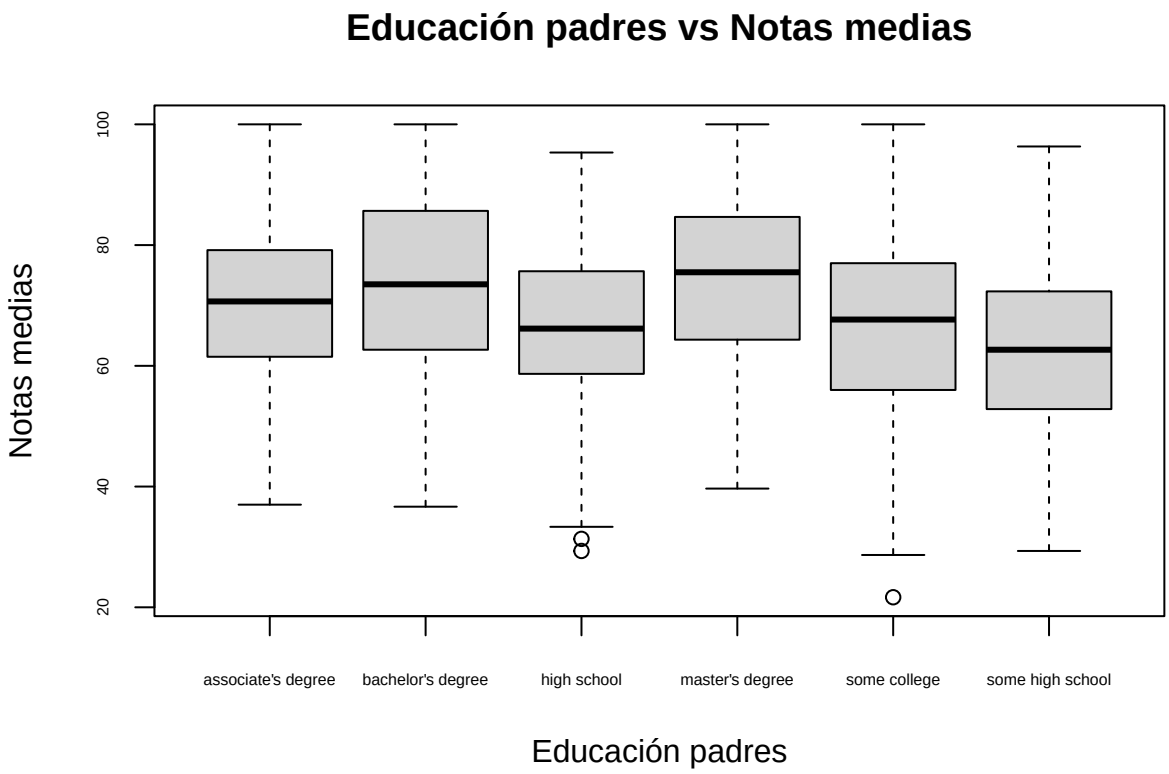
Las notas de amos géneros son muy similares en general, aunque muy ligeramente superiores las de las mujeres.

```
plot(data$race.ethnicity, data$meangrade, main = 'Raza vs Notas medias', ylab = 'Notas
↪  medias', xlab = 'Raza')
```



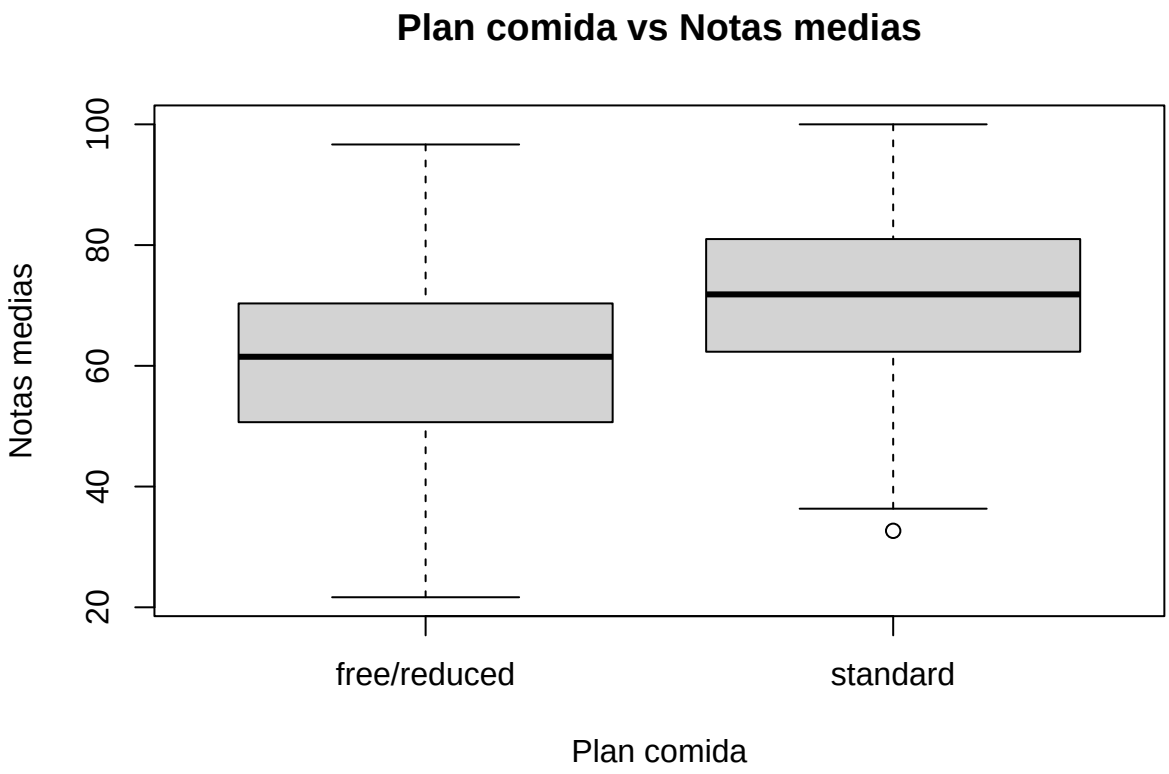
La raza mas frecuente (C) también es la que obtiene peores resultados académicos.

```
plot(data$parental.level.of.education, data$meangrade, cex.axis = 0.5, main = 'Educación
↪  padres vs Notas medias', ylab = 'Notas medias', xlab = 'Educación padres')
```



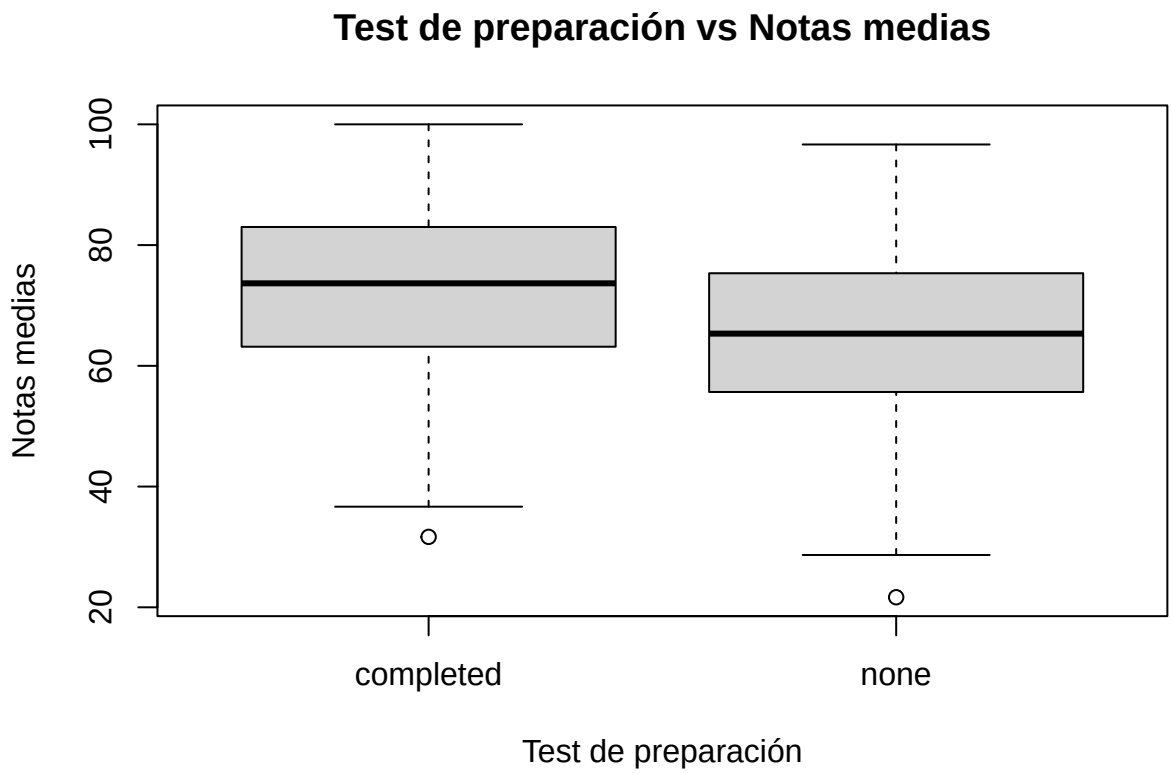
Los resultados son esperados, los niños con padres estudiosos tienen mejores notas.

```
plot(data$lunch, data$meangrade, main = 'Plan comida vs Notas medias', ylab = 'Notas
↪  medias', xlab = 'Plan comida')
```



Los estudiantes con el plan de comida estándar tienen mejores notas.

```
plot(data$test.preparation.course, data$meangrade, main = 'Test de preparación vs Notas
↪  medias', ylab = 'Notas medias', xlab = 'Test de preparación')
```

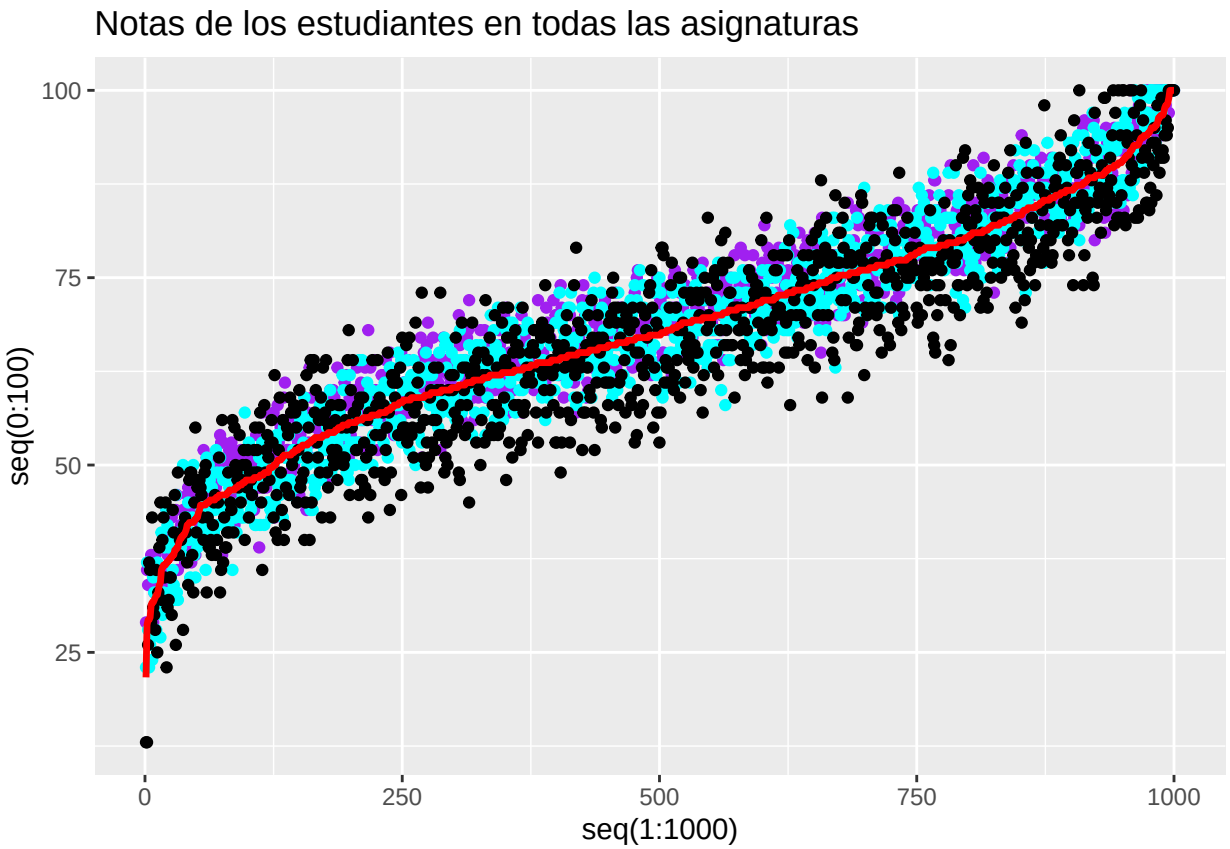


Y los estudiantes con el test de preparación tienen mejores resultados.

Veamos ahora si por lo general los estudiantes suelen sacar notas parecidas en todas las asignaturas.

```
newdata <- data[order(data$meangrade),]

ggplot(newdata, aes(x=seq(1:1000), y=seq(0:100)))+
  geom_point(data = newdata, aes(x = seq(1:1000), y = reading.score), colour = 'purple')+
  geom_point(data = newdata, aes(x = seq(1:1000), y = writing.score), colour = 'cyan')+
  geom_point(data = newdata, aes(x = seq(1:1000), y = math.score), colour = 'black')+
  geom_line(data = newdata, aes(x = seq(1:1000), y = meangrade), colour = 'red', size =
  ↪  1.2)+
  ggtitle("Notas de los estudiantes en todas las asignaturas")
```



Como podemos ver hay una clara linea de tendencia, siendo los puntos negros las notas de los estudiantes en matemáticas, los morados las notas de lectura, los cían las notas de escritura y la línea roja la media. También podemos apreciar que los estudiantes parecen tener notas más bajas en matemáticas respecto a las otras asignaturas.

Resolución del problema

Tras el análisis realizado, podemos concluir que no hay diferencia significativas en las notas entre mujeres y hombres. Pero sí hay diferencia en relación a la raza y la educación de los padres, donde los grupos D y E sobresalen de los demás y los hijos con padres con estudios superiores también reciben mejores notas. Además, los estudiantes con el plan de comida estándar tiene mejores notas, esto puede significar que los estudiantes con un nivel socio-económico mas alto tienen mejores resultados académicos.

Por último, cabe destacar que los estudiantes tienen tendencia a sacar notas similares en las asignaturas mostradas.

Versiones

```
sessionInfo()
```

```
## R version 4.2.1 (2022-06-23 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
```

```
## Running under: Windows 10 x64 (build 22621)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United Kingdom.utf8
## [2] LC_CTYPE=English_United Kingdom.utf8
## [3] LC_MONETARY=English_United Kingdom.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United Kingdom.utf8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ggplot2_3.3.6
##
## loaded via a namespace (and not attached):
##  [1] pillar_1.8.1     compiler_4.2.1   highr_0.9        tools_4.2.1
##  [5] digest_0.6.29    evaluate_0.17    lifecycle_1.0.3  tibble_3.1.8
##  [9] gtable_0.3.1     pkgconfig_2.0.3  rlang_1.0.6      cli_3.4.1
## [13] DBI_1.1.3        rstudioapi_0.14  yaml_2.3.5       xfun_0.33
## [17] fastmap_1.1.0    withr_2.5.0      stringr_1.4.1    dplyr_1.0.10
## [21] knitr_1.40       generics_0.1.3   vctrs_0.5.1      grid_4.2.1
## [25] tidyselect_1.2.0 glue_1.6.2       R6_2.5.1         fansi_1.0.3
## [29] rmarkdown_2.17   farver_2.1.1     magrittr_2.0.3   scales_1.2.1
## [33] htmltools_0.5.3  assertthat_0.2.1 colorspace_2.0-3 labeling_0.4.2
## [37] utf8_1.2.2       stringi_1.7.8    munsell_0.5.0
```