# Tipología y ciclo de vida de los datos

## Práctica 2: Limpieza y validación de los datos

**Iosu Rodríguez Alfaro**

# Índice

1. Descripción del dataset ..... 3
2. Integración y selección de los datos de interés a analizar ..... 4
3. Limpieza de los datos ..... 5
3.1. Valores nulos o ceros ..... 5
3.2.Valores extremos o outliers ..... 5
4. Análisis de los datos ..... 8
4.1. Selección de los grupos de datos a analizar ..... 8
4.2. Comprobación de la normalidad y homogeneidad de la varianza ..... 9
Normalización de los datos ..... 9
4.3. Aplicación de pruebas estadísticas ..... 10
4.3.1.¿Qué factores influyen más a la hora de que un alumno saque mejores notas? ..... 10
4.3.2.¿En que asignaturas son mejores los hombres y en cuales las mujeres? ..... 13
4.3.3. ¿La nota media de los alumnos es mayor para aquellos con padres con estudios universitarios? ..... 13
4.3.4. ¿Podríamos saber a partir de los datos de los alumnos, antes de realizar los exámenes, que previsión hay de aprobados y suspensos en función de sus datos? ..... 14
5. Conclusiones ..... 20

# 1. Descripción del dataset

El conjunto de datos con el que se trabaja se ha obtenido del siguiente enlace de Kaggle:

https://www.kaggle.com/spscientist/students-performance-in-exams

El dataset se llama "Students Performance in Exams" y contiene las notas de 1000 alumnos de educación secundaria de Estados Unidos en varias asignaturas así como otra información.

El dataset tiene las siguientes columnas (8):

- Gender: contiene el género del alumno (female o male)
- Race/ethnicity: indica la raza/etnia del alumno. Los valores posibles son group A, group B, group C, group D, group E.
- Parental level of education: indica el nivel de estudios de los padres de los alumnos.
- Lunch: indica el tipo de comida que toma el alumno (estándar o reducido/gratis)
- Test preparation curse: indica si el alumno ha realizado el curso de preparación de exámenes.
- Math score: nota del alumno en matemáticas.
- Reading score: nota en lectura del alumno.
- Writing score: nota en escritura del alumno.

El dataset es importante ya que permite estudiar la relación que hay entre las notas obtenidas por los alumnos y diversos factores como la raza, género, nivel de educación de los padres…

Las preguntas que se pretenden responder son:

- ¿Qué factores influyen más a la hora de que un alumno saque mejores notas?
- ¿En qué asignaturas son mejores los hombres y en cuales las mujeres?
- ¿La nota media de los alumnos es mayor para aquellos con padres con estudios universitarios?
- ¿Podríamos saber a partir de los datos de los alumnos, antes de realizar los exámenes, que previsión hay de aprobados y suspensos en función de sus datos?

# 2. Integración y selección de los datos de interés a analizar

El conjunto de datos está contenido en un único fichero csv, por lo que no hay necesidad de integrar datos. Sí que se va a realizar una limpieza de los datos y se van a generar nuevas variables de interés. Se van a mantener todas las variables del dataset, pero las notas las tendremos en una escala entre 0 y 1, dividiendo las del conjunto de datos entre 100 (están en una escala de 0 a 100). También generaremos dos nuevas variables, una con la nota media de las tres asignaturas y otra para decir si la media está aprobada (>=0.5) o suspensa (<0.5). También vamos a reducir las categorías de niveles de estudio de los padres. Aquellos que su nivel de estudio sea “high school” o “some high school” se considerarán que no tienen estudios universitarios y el resto que si los tienen. Con estos datos y variables generaremos un nuevo csv.

# 3. Limpieza de los datos

## 3.1. Valores nulos o ceros

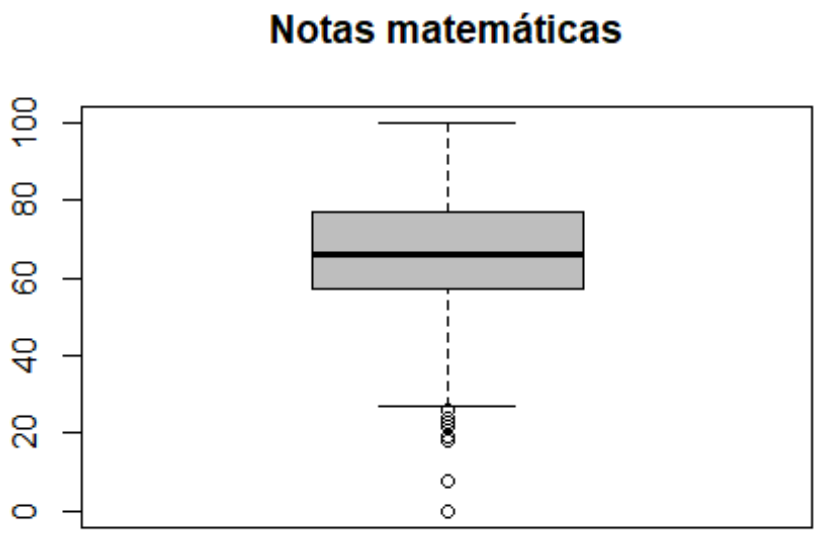
```
# Mostramos cuantos valores son nulos o 0
sapply(data, function(x) sum(is.na(x) || x == 0))
```

```
##                       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
```

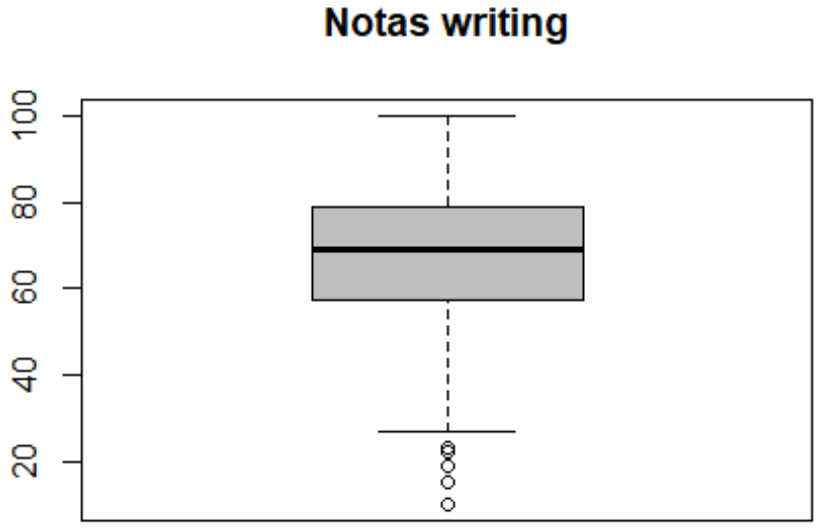
Observamos que ningún campo tiene valores nulos o 0.

## 3.2.Valores extremos o outliers

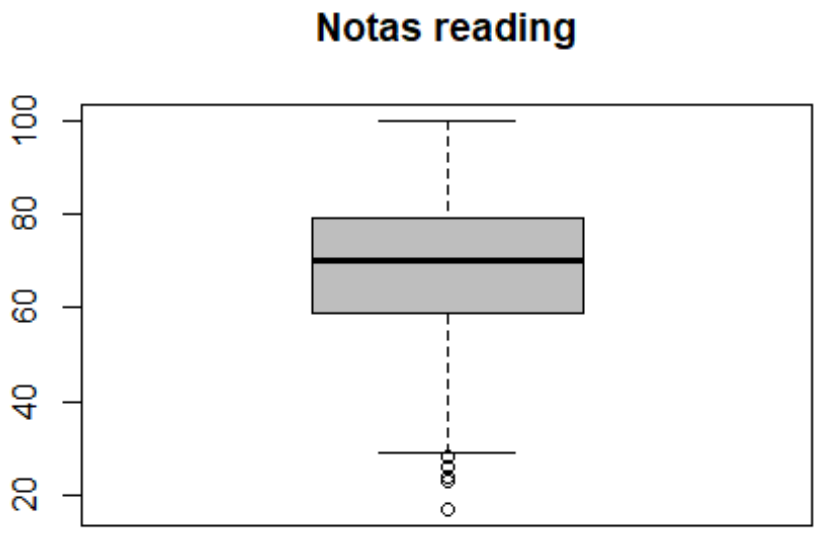
```
# Mostramos un boxplot de cada variable numérica para observar si hay val
ores extremos.
boxplot(data$math.score,main="Notas matemáticas",col="gray")
```




```r
boxplot(data$writing.score,main="Notas writing",col="gray")
```




```r
boxplot(data$reading.score,main="Notas reading",col="gray")
```




Se observan varios valores extremos por debajo en las tres asignaturas. Sin embargo no nos preocupa. Se deben a alumnos que han sacado una nota muy baja, pero ya que no hay notas que superen la máxima posible (100) o sean menor que la mínima posible (<0), los mantenemos.

# 4. Análisis de los datos

## 4.1. Selección de los grupos de datos a analizar

Además de los datos que disponemos, queremos añadir dos nuevos campos. Uno para indicar la media de las tres asignaturas de cada alumno y otro para indicar si con su media supera el aprobado o no. Se considera aprobado, una nota media igual o mayor a 50 (posteriormente se reducirá la escala a 0-1).

También vamos a unificar los niveles de estudio de los padres en 2. Aquellos con estudios universitarios (university) y aquellos que no (no_university). Se considera que no tienen estudios universitarios aquellos con valor 'high school', 'some high school'.

```
# Añadimos la variable media
data$mean.score = as.numeric((data$math.score + data$reading.score + data
$writing.score)/3)
# Añadimos la variable aprobado
data$aprobado = as.integer(data$mean.score >=50)

#Unificamos los niveles de educación de los padres
data$parental.level.of.education = as.factor(gsub("high school|some high
school", "no_university",data$parental.level.of.education, ignore.case =
TRUE))
data$parental.level.of.education<-as.character(data$parental.level.of.edu
cation)
data$parental.level.of.education = as.factor(replace(data$parental.level.
of.education, data$parental.level.of.education != "no_university", "unive
rsity"))

summary(data$parental.level.of.education)

## no_university    university
##           375           625

sapply(data,class)

##                       gender              race.ethnicity
##                     "factor"                    "factor"
## parental.level.of.education                       lunch
##                     "factor"                    "factor"
##     test.preparation.course                  math.score
##                     "factor"                   "integer"
##                reading.score                 writing.score
##                    "integer"                   "integer"
##                   mean.score                    aprobado
##                    "numeric"                   "integer"
```

## 4.2. Comprobación de la normalidad y homogeneidad de la varianza

### Normalización de los datos

Vamos a aplicar el test de Shapiro Wilk en cada variable numérica para ver si está normalizada.

```
shapiro.test(data$math.score)

##
##  Shapiro-Wilk normality test
##
## data:  data$math.score
## W = 0.99315, p-value = 0.0001455

shapiro.test(data$reading.score)

##
##  Shapiro-Wilk normality test
##
## data:  data$reading.score
## W = 0.99292, p-value = 0.0001055

shapiro.test(data$writing.score)

##
##  Shapiro-Wilk normality test
##
## data:  data$writing.score
## W = 0.99196, p-value = 2.922e-05

shapiro.test(data$mean.score)

##
##  Shapiro-Wilk normality test
##
## data:  data$mean.score
## W = 0.99315, p-value = 0.0001453
```

En este caso, las variables numéricas son valores entre 0 y 100 y todas están normalizadas puesto que su p-valor es < 0.05. Sin embargo vamos a hacer que estos valores queden entre 0 y 1. Tan solo dividiremos entre 100.

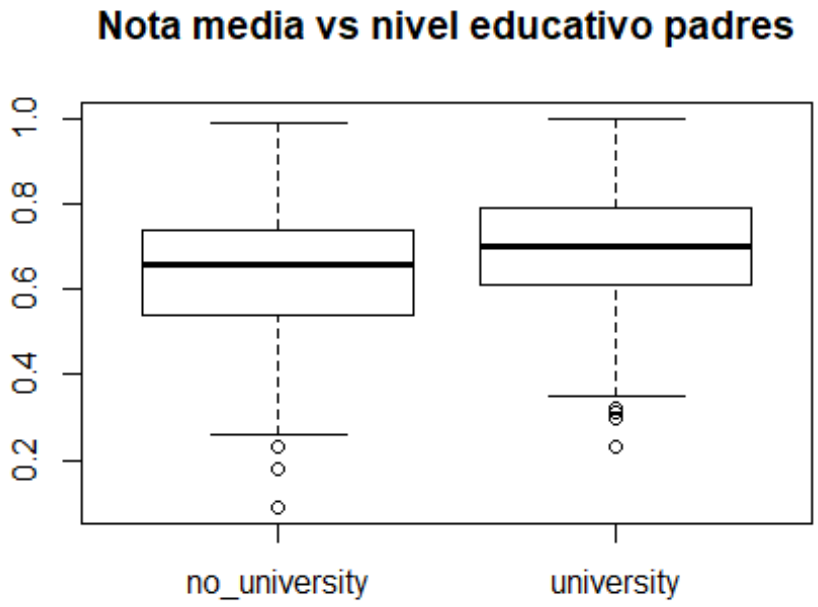
```
max = 100
data$math.score = as.numeric(data$math.score/max)
data$reading.score = as.numeric(data$reading.score/max)
data$writing.score = as.numeric(data$writing.score/max)
data$mean.score = as.numeric(round(data$mean.score/max, 2))
```

## 4.3. Aplicación de pruebas estadísticas

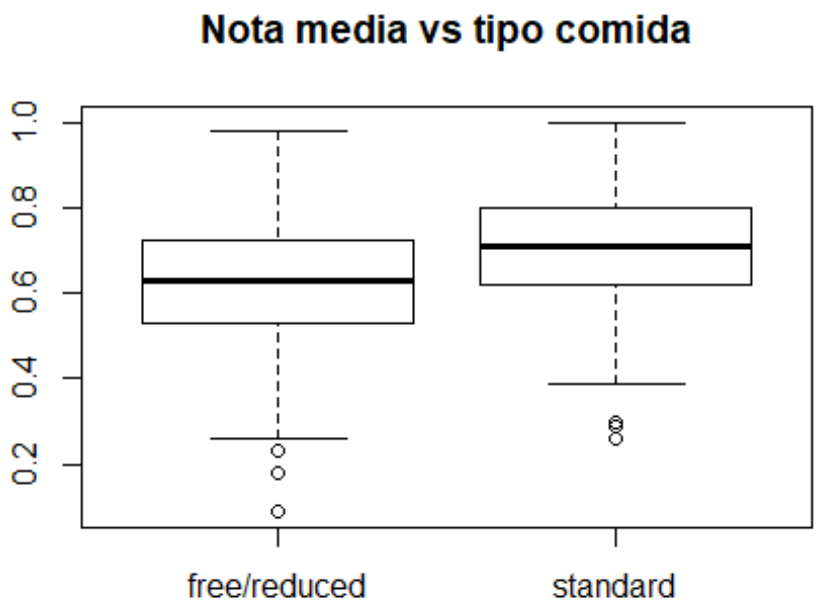
### 4.3.1.¿Qué factores influyen más a la hora de que un alumno saque mejores notas?

Para responder a esta pregunta vamos a observar la relación entre la nota media y cada variable categórica.

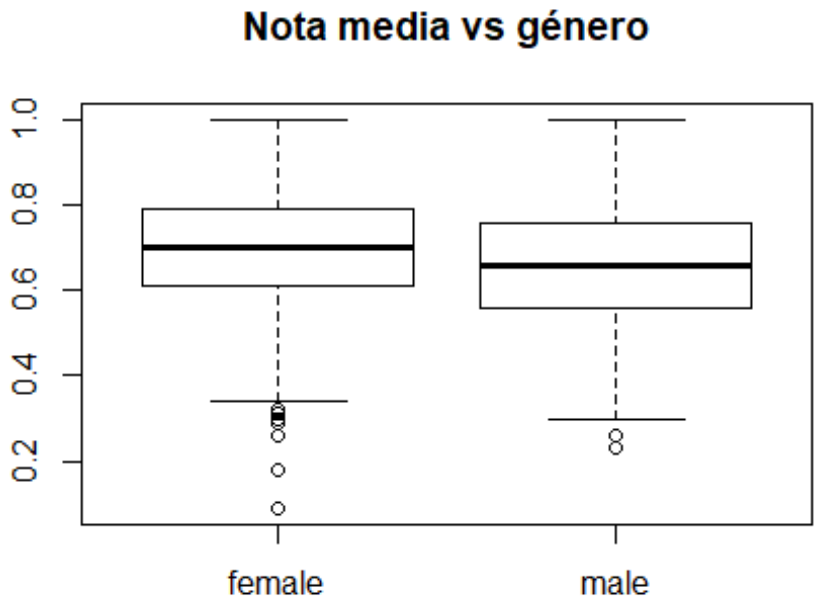
```
boxplot(data$mean.score ~ data$parental.level.of.education,main="Nota med
ia vs nivel educativo padres")
```



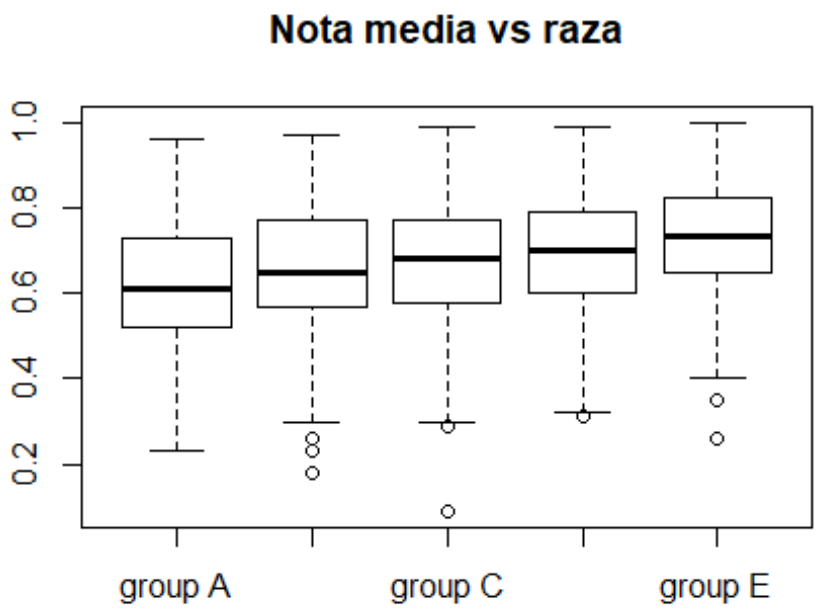

```
boxplot(data$mean.score ~ data$lunch,main="Nota media vs tipo comida")
```



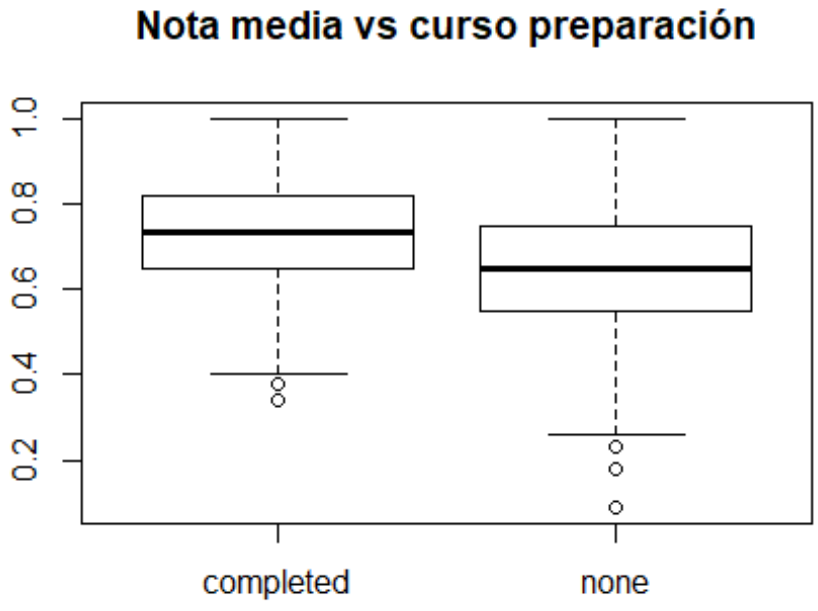

```r
boxplot(data$mean.score ~ data$gender,main="Nota media vs género")
```




```r
boxplot(data$mean.score ~ data$race.ethnicity,main="Nota media vs raza")
```



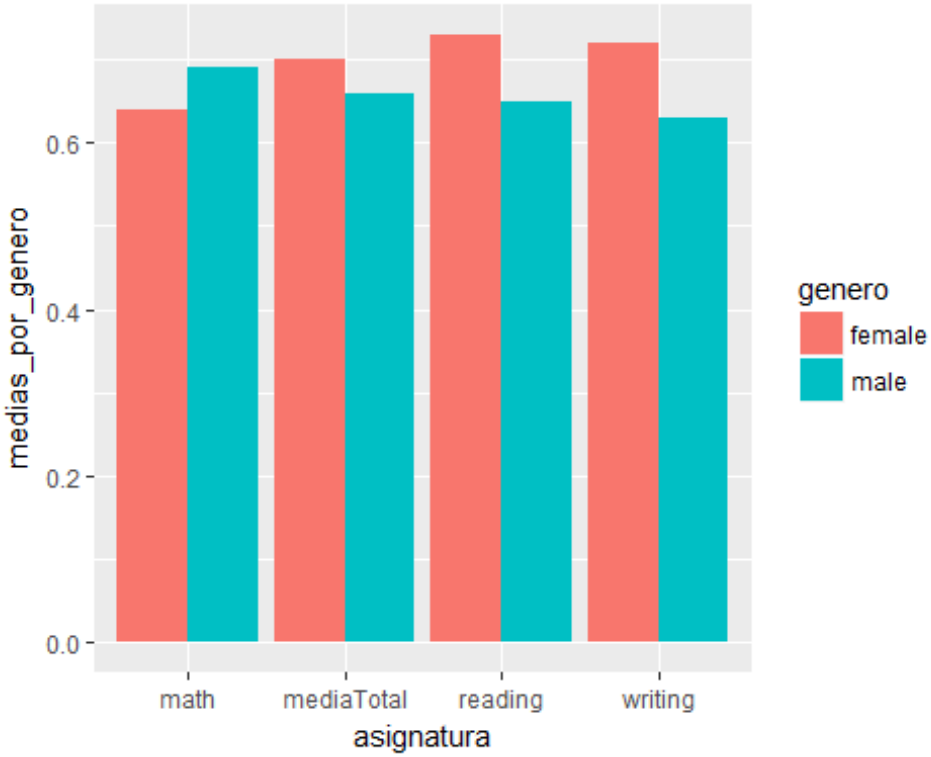

```r
boxplot(data$mean.score ~ data$test.preparation.course,main="Nota media v
s curso preparación")
```



Viendo el nivel de solapamiento que hay, podemos decir que la raza, y el haber realizado el curso de preparación son los factores que más influyen en la nota media. El sexo apenas tiene influencia y el nivel educativo de los padres y el tipo de comida, tiene cierta influencia.

### 4.3.2.¿En que asignaturas son mejores los hombres y en cuales las mujeres?

Obtenemos una gráfica con las medias de cada asignatura y totales por género:




Podemos asegurar que las mujeres han obtenido mejor media. Que los hombres han sacado mejor nota media en matemáticas y las mujeres en Reading y writing.

### 4.3.3. ¿La nota media de los alumnos es mayor para aquellos con padres con estudios universitarios?

Para ello vamos a realizar un contraste de dos muestras sobre la diferencia de las media. Sabemos que los datos siguen una distribución normal. Aun así, como el número de muestras es grande, por el teorema del límite central, podemos considerar que sigue una distribución normal.

Planteamos las hipótesis: -Hipótesis nula: H0: mu1-mu2=0. La nota media es la misma tanto si los padres tienen estudios universitarios (mu1) como si no (m2).

-Hipótesis alternativa: H1: mu1-mu2>0 la nota media es superior si los padres tienen estudios universitarios.

```
padresuniv.si<-data[ data$parental.level.of.education == 'university', ]
padresuniv.no<-data[ data$parental.level.of.education == 'no_university',
```

```
]
t.test(padresuniv.si$mean.score, padresuniv.no$mean.score, alternative="g
reater", conf.level=0.99, paired=FALSE, var.equal=FALSE)

##
##  Welch Two Sample t-test
##
## data:  padresuniv.si$mean.score and padresuniv.no$mean.score
## t = 6.4502, df = 767.89, p-value = 9.878e-11
## alternative hypothesis: true difference in means is greater than 0
## 99 percent confidence interval:
##  0.03789259        Inf
## sample estimates:
## mean of x mean of y
## 0.6998720 0.6405333
```

Puesto que el p-valor es 9.878e-11 < 0.01, rechazamos la hipótesis nula y aceptamos la hipótesis alternativa. Por tanto podemos decir con un 99% de confianza que la nota media de los alumnos es mayor en aquellos que tienen padres universitarios.

### 4.3.4. ¿Podríamos saber a partir de los datos de los alumnos, antes de realizar los exámenes, que previsión hay de aprobados y suspensos en función de sus datos?

Vamos a aplicar la regresión logística considerando como variables todos los factores. De esta forma crearemos un modelo para clasificar a los alumnos como aprobados o suspendidos en función de sus datos.

```
model = glm(data$aprobado~data$gender+data$parental.level.of.education+da
ta$lunch+data$race.ethnicity+data$test.preparation.course, family=binomia
l(link='logit'))
summary(model)

##
## Call:
## glm(formula = data$aprobado ~ data$gender + data$parental.level.of.edu
cation +
##     data$lunch + data$race.ethnicity + data$test.preparation.course,
##     family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6879   0.2143   0.3233   0.4788   1.2304
##
## Coefficients:
##                                        Estimate Std. Error z value
## (Intercept)                             1.75615    0.41818   4.200
## data$gendermale                        -0.72332    0.22731  -3.182
```

```
## data$parental.level.of.educationuniversity  1.06968    0.22416   4.772
## data$lunchstandard                          1.20540    0.22414   5.378
## data$race.ethnicitygroup B                  0.05033    0.37239   0.135
## data$race.ethnicitygroup C                  0.43921    0.35806   1.227
## data$race.ethnicitygroup D                  0.88752    0.38800   2.287
## data$race.ethnicitygroup E                  0.71055    0.46626   1.524
## data$test.preparation.coursenone           -1.15659    0.27799  -4.161
##                                            Pr(>|z|)
## (Intercept)                                2.67e-05 ***
## data$gendermale                             0.00146 **
## data$parental.level.of.educationuniversity 1.82e-06 ***
## data$lunchstandard                         7.54e-08 ***
## data$race.ethnicitygroup B                  0.89248
## data$race.ethnicitygroup C                  0.21996
## data$race.ethnicitygroup D                  0.02217 *
## data$race.ethnicitygroup E                  0.12753
## data$test.preparation.coursenone           3.17e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 663.25  on 999  degrees of freedom
## Residual deviance: 571.31  on 991  degrees of freedom
## AIC: 589.31
##
## Number of Fisher Scoring iterations: 6
```

Observamos que el modelo obtenido no es demasiado bueno. Posiblemente necesitemos más datos, o simplemente, clasificar a los alumnos por sus factores como posibles aprobados o suspendidos, no es factible, ya que las posibles combinaciones de los valores que toman los factores son limitadas. Veremos con este modelo cuantos acertamos y fallamos.

```
prob_acc <-predict(model,data, type = "response")
pred_acc <-ifelse (prob_acc>= 0.5, 1, 0)
table(data$aprobado,pred_acc)
```

```
##    pred_acc
##       0   1
##   0   8  95
##   1   6 891
```

Con este modelo vemos que somo muy optimistas, dando por aprobados a la gran mayoría de suspendidos. Sin embargo, si predecimos bien aquellos que han aprobado.

Ahora vamos a ver si conseguimos un modelo que sirva para predecir los aprobados y suspensos de cada asignatura en función de los factores.

```
accMath = as.integer(data$math.score >=0.5)
accReading = as.integer(data$reading.score >=0.5)
accWriting = as.integer(data$writing.score >=0.5)

modelMath = glm(accMath~data$gender+data$parental.level.of.education+data
$lunch+data$race.ethnicity+data$test.preparation.course, family=binomial(
link='logit'))
summary(modelMath )

##
## Call:
## glm(formula = accMath ~ data$gender + data$parental.level.of.education
+
##     data$lunch + data$race.ethnicity + data$test.preparation.course,
##     family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6254   0.2219   0.3471   0.5210   1.4046
##
## Coefficients:
##                                          Estimate Std. Error z value
## (Intercept)                                0.4598     0.3741   1.229
## data$gendermale                            0.6349     0.2078   3.055
## data$parental.level.of.educationuniversity   0.7015     0.2028   3.458
## data$lunchstandard                         1.8300     0.2113   8.659
## data$race.ethnicitygroup B                 0.2836     0.3581   0.792
## data$race.ethnicitygroup C                 0.4664     0.3383   1.379
## data$race.ethnicitygroup D                 0.7675     0.3569   2.150
## data$race.ethnicitygroup E                 1.0840     0.4549   2.383
## data$test.preparation.coursenone          -0.9796     0.2377  -4.121
##                                          Pr(>|z|)
## (Intercept)                              0.219073
## data$gendermale                          0.002250 **
## data$parental.level.of.educationuniversity 0.000544 ***
## data$lunchstandard                        < 2e-16 ***
## data$race.ethnicitygroup B               0.428499
## data$race.ethnicitygroup C               0.168002
## data$race.ethnicitygroup D               0.031522 *
## data$race.ethnicitygroup E               0.017169 *
## data$test.preparation.coursenone         3.77e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 791.56  on 999  degrees of freedom
## Residual deviance: 657.76  on 991  degrees of freedom
## AIC: 675.76
##
## Number of Fisher Scoring iterations: 5
```

```
modelReading = glm(accReading~data$gender+data$parental.level.of.educatio
n+data$lunch+data$race.ethnicity+data$test.preparation.course, family=bin
omial(link='logit'))
summary(modelReading)
```

```
##
## Call:
## glm(formula = accReading ~ data$gender + data$parental.level.of.educat
ion +
##     data$lunch + data$race.ethnicity + data$test.preparation.course,
##     family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7111   0.2234   0.3339   0.4825   1.1341
##
## Coefficients:
##                                          Estimate Std. Error z value
## (Intercept)                                2.1518     0.4394   4.897
## data$gendermale                           -0.9579     0.2447  -3.915
## data$parental.level.of.educationuniversity   0.7563     0.2327   3.249
## data$lunchstandard                         1.0419     0.2336   4.460
## data$race.ethnicitygroup B                 0.1587     0.3820   0.415
## data$race.ethnicitygroup C                 0.6431     0.3714   1.732
## data$race.ethnicitygroup D                 0.8197     0.3894   2.105
## data$race.ethnicitygroup E                 0.7904     0.4813   1.642
## data$test.preparation.coursenone          -1.0910     0.2917  -3.740
##                                          Pr(>|z|)
## (Intercept)                              9.71e-07 ***
## data$gendermale                          9.04e-05 ***
## data$parental.level.of.educationuniversity 0.001157 **
## data$lunchstandard                       8.22e-06 ***
## data$race.ethnicitygroup B               0.677869
## data$race.ethnicitygroup C               0.083328 .
## data$race.ethnicitygroup D               0.035293 *
## data$race.ethnicitygroup E               0.100505
## data$test.preparation.coursenone         0.000184 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 605.08  on 999  degrees of freedom
## Residual deviance: 533.88  on 991  degrees of freedom
```

```
## AIC: 551.88
##
## Number of Fisher Scoring iterations: 6
```

```
modelWriting = glm(accWriting~data$gender+data$parental.level.of.educatio
n+data$lunch+data$race.ethnicity+data$test.preparation.course, family=bin
omial(link='logit'))
summary(modelWriting)
```

```
##
## Call:
## glm(formula = accWriting ~ data$gender + data$parental.level.of.educat
ion +
##     data$lunch + data$race.ethnicity + data$test.preparation.course,
##     family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6634   0.2000   0.3475   0.4971   1.4145
##
## Coefficients:
##                                          Estimate Std. Error z value
## (Intercept)                                2.2529     0.4231   5.324
## data$gendermale                           -1.3061     0.2342  -5.576
## data$parental.level.of.educationuniversity  0.9422     0.2165   4.351
## data$lunchstandard                         0.9869     0.2171   4.545
## data$race.ethnicitygroup B                 0.2778     0.3580   0.776
## data$race.ethnicitygroup C                 0.7393     0.3443   2.147
## data$race.ethnicitygroup D                 1.0261     0.3639   2.820
## data$race.ethnicitygroup E                 0.7839     0.4327   1.811
## data$test.preparation.coursenone          -1.4888     0.2892  -5.148
##                                          Pr(>|z|)
## (Intercept)                              1.01e-07 ***
## data$gendermale                          2.46e-08 ***
## data$parental.level.of.educationuniversity 1.35e-05 ***
## data$lunchstandard                       5.49e-06 ***
## data$race.ethnicitygroup B                0.43774
## data$race.ethnicitygroup C                0.03179 *
## data$race.ethnicitygroup D                0.00481 **
## data$race.ethnicitygroup E                0.07007 .
## data$test.preparation.coursenone         2.63e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 709.59  on 999  degrees of freedom
## Residual deviance: 593.38  on 991  degrees of freedom
## AIC: 611.38
```

```
##
## Number of Fisher Scoring iterations: 6
```

```
print('Predicciones para matemáticas')
```

```
## [1] "Predicciones para matemáticas"
```

```
prob_acc <-predict(modelMath,data, type = "response")
pred_acc <-ifelse (prob_acc>= 0.5, 1, 0)
table(accMath,pred_acc)
```

```
##        pred_acc
## accMath   0   1
##       0  21 114
##       1  11 854
```

```
print('Predicciones para reading')
```

```
## [1] "Predicciones para reading"
```

```
prob_acc <-predict(modelReading,data, type = "response")
pred_acc <-ifelse (prob_acc>= 0.5, 1, 0)
table(accReading,pred_acc)
```

```
##           pred_acc
## accReading   1
##          0  90
##          1 910
```

```
print('Predicciones para writing')
```

```
## [1] "Predicciones para writing"
```

```
prob_acc <-predict(modelWriting,data, type = "response")
pred_acc <-ifelse (prob_acc>= 0.5, 1, 0)
table(accWriting,pred_acc)
```

```
##           pred_acc
## accWriting   0   1
##          0  10 104
##          1   4 882
```

Como observamos seguimos sin poder obtener un modelo bueno para poder clasificar con cierta seguridad los aprobados y suspensos a partir de los datos de los alumnos, antes de que realicen los exámenes. Concluimos que con los factores de los alumnos no somo capaces de clasificar a estos como candidatos a aprobar o suspender.

# 5. Conclusiones

Podemos concluir por tanto lo siguiente:

- Los factores que más influyen en la nota media de los alumnos son la raza y el hecho de haber realizado curso de preparación. El nivel educativo de los padres y el tipo de comida, son factores que tienen cierta influencia. El sexo no es considerado como un factor muy influyente, aunque si se observan diferencias entre ambos sexos.

- Los hombres, de media, tienen mejor nota que las mujeres en matemáticas, mientras que las mujeres tienen mejor media global y mejor media en reading y writing.

- Podemos asegurar con un 99% de confianza que la nota media de los alumnos es mayor en aquellos que tienen padres con estudios universitarios.

- A partir de los factores de los alumnos, no podemos realizar una previsión fiable de si los alumnos aprobaran o suspenderán cada asignatura, antes de realizar los exámenes. Esto es debido a que todas las variables empleadas son las categóricas y que la combinación de los valores que estas pueden tomar es limitado.