

# Tipología y ciclo de vida de los datos

**Alumnos: Eduardo Ranedo Martínez / Luis Piñuela Galán**

**Fecha: 14/06/2023**

**PRAC: PRAC 2**

## Índice:

1. Descripción del dataset
2. Integración y selección
3. Limpieza de los datos
4. Análisis de los datos
5. Representación de los resultados
6. Resolución del problema
7. Código
8. Vídeo

| Contribuciones            | Firma  |
|---------------------------|--|
| Investigación previa      | Eduardo Ranedo Martínez / Luis Piñuela Galán |
| Redacción de respuestas   | Eduardo Ranedo Martínez / Luis Piñuela Galán |
| Desarrollo de código      | Eduardo Ranedo Martínez / Luis Piñuela Galán |
| Participación en el video | Eduardo Ranedo Martínez / Luis Piñuela Galán |

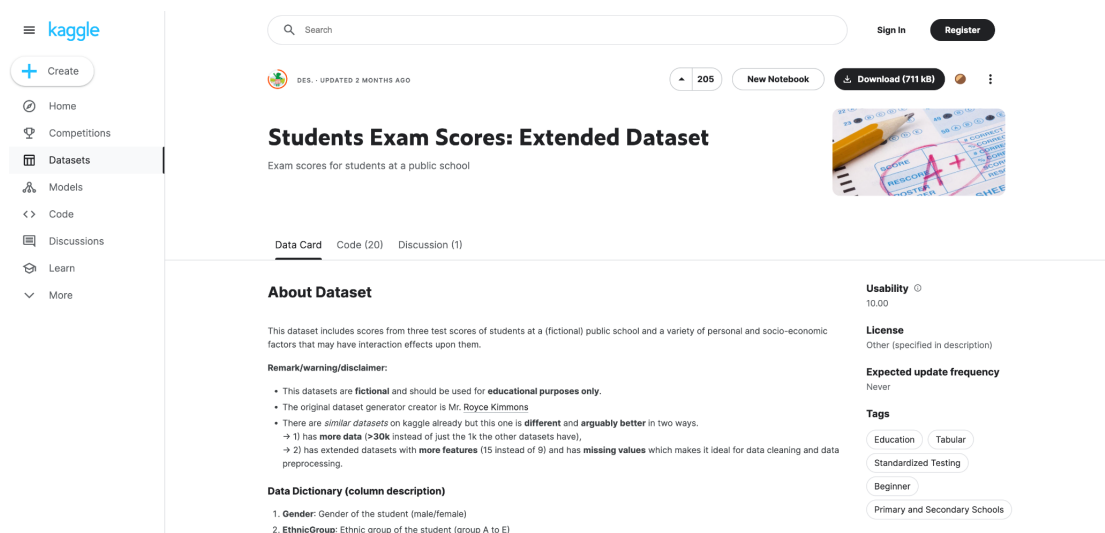
## 1. Descripción del dataset

El dataset escogido nos muestra información sobre el rendimiento académico de diferentes estudiantes con un total de 15 variables que nos pueden ayudar a realizar el análisis desde diferentes perspectivas. Por ejemplo, se podría evaluar el rendimiento académico en función del sexo o el origen étnico, identificando posibles brechas o desigualdades. Asimismo, se podría analizar la relación entre el rendimiento académico y el nivel educativo de los padres, explorando el impacto de este factor en los resultados obtenidos.

Nuestro dataset contiene más de 30.000 registros, lo que proporciona una muestra lo suficientemente grande como para obtener conclusiones estadísticamente significativas y representativas de la población en estudio. Esto es especialmente relevante cuando se trata de datos académicos, ya que permite analizar patrones y tendencias en el rendimiento académico de una amplia variedad de individuos.

El análisis del rendimiento académico puede ser de interés para instituciones educativas, investigadores, padres y estudiantes. Los resultados obtenidos a partir de este dataset podrían utilizarse para identificar patrones de éxito académico, proponer estrategias de mejora, o incluso para informar políticas educativas.

**Dataset:** <https://www.kaggle.com/datasets/desalegngeb/students-exam-scores>



The screenshot shows the Kaggle dataset page for 'Students Exam Scores: Extended Dataset'. The page includes a search bar, navigation links (Home, Competitions, Datasets, Models, Code, Discussions, Learn, More), and a sidebar with a 'Create' button. The main content area displays the dataset title, a brief description ('Exam scores for students at a public school'), and a 'Data Card' tab. The 'About Dataset' section provides details about the data source, a disclaimer, and a data dictionary. The 'Usability' section shows a score of 10.00, and the 'License' section indicates 'Other (specified in description)'. The 'Expected update frequency' is 'Never', and the 'Tags' include 'Education', 'Tabular', 'Standardized Testing', 'Beginner', and 'Primary and Secondary Schools'.

## 2. Integración y selección

Para nuestro estudio, hemos considerado que el dataset ya contiene suficiente información para realizar los análisis planteados. Sin embargo, sí que hemos hecho una selección de las variables que vamos a utilizar, ya que hay algunas que no nos aportan valor. En este sentido hemos pasado de un total de 15 variables a 12. Se han eliminado las variables del ID que identifica a los individuos, la variable TestPre y LunchType.

### 3. Limpieza de los datos

- a. **Eliminación de datos innecesarios:** eliminamos del análisis las columnas que a nuestro juicio de análisis no aportan información relevante, como X y Lunch Type

```
# Eliminamos columnas innecesarias
dataset <- subset(data, select = -c(X, LunchType))
str(dataset)
...

'data.frame': 30641 obs. of 14 variables:
 $ Gender      : chr  "female" "female" "female" "male" ...
 $ EthnicGroup : chr  "" "group C" "group B" "group A" ...
 $ ParentEduc  : chr  "bachelor's degree" "some college" "master's degree" "associate's degree" ...
 $ LunchType   : chr  "standard" "standard" "standard" "free/reduced" ...
 $ ParentMaritalStatus: chr  "married" "married" "single" "married" ...
 $ PracticeSport : chr  "regularly" "sometimes" "sometimes" "never" ...
 $ IsFirstChild : chr  "yes" "yes" "yes" "no" ...
 $ NrSiblings  : int   3 0 4 1 0 1 1 1 3 NA ...
 $ TransportMeans : chr  "school_bus" "" "school_bus" "" ...
 $ WklyStudyHours : chr  "< 5" "5-oct" "< 5" "5-oct" ...
 $ MathScore   : int   71 69 87 45 76 73 85 41 65 37 ...
 $ ReadingScore : int   71 90 93 56 78 84 93 43 64 59 ...
 $ WritingScore : int   74 88 91 42 75 79 89 39 68 50 ...
 $ X           : logi  NA NA NA NA NA NA ...

'data.frame': 30641 obs. of 12 variables:
 $ Gender      : chr  "female" "female" "female" "male" ...
 $ EthnicGroup : chr  "" "group C" "group B" "group A" ...
 $ ParentEduc  : chr  "bachelor's degree" "some college" "master's degree" "associate's degree" ...
 $ ParentMaritalStatus: chr  "married" "married" "single" "married" ...
 $ PracticeSport : chr  "regularly" "sometimes" "sometimes" "never" ...
 $ IsFirstChild : chr  "yes" "yes" "yes" "no" ...
 $ NrSiblings  : int   3 0 4 1 0 1 1 1 3 NA ...
 $ TransportMeans : chr  "school_bus" "" "school_bus" "" ...
 $ WklyStudyHours : chr  "< 5" "5-oct" "< 5" "5-oct" ...
 $ MathScore   : int   71 69 87 45 76 73 85 41 65 37 ...
 $ ReadingScore : int   71 90 93 56 78 84 93 43 64 59 ...
 $ WritingScore : int   74 88 91 42 75 79 89 39 68 50 ...
```

- b. **Tratamiento de valores nulos:** Se ha realizado una búsqueda de las dimensiones que puedan contener valores nulos, se ha decidido que dado la gran cantidad de registros se va a optar a su eliminación para no enturbiar el análisis. Esto ha reducido el número de registros número de registros de 30641 a 20445.

|   |                     |                    |                             |                      |                     |                    |                        |
|---|---------------------|--------------------|-----------------------------|----------------------|---------------------|--------------------|------------------------|
| <pre>## {r} colSums(is.na(dataset)) ##</pre>  |                     |                    |                             |                      |                     |                    |                        |
| Gender<br>0                                   | EthnicGroup<br>0    | ParentEduc<br>0    | ParentMaritalStatus<br>0    | PracticeSport<br>0   | IsFirstChild<br>0   | NrSiblings<br>1572 | TransportMeans<br>0    |
| WklyStudyHours<br>0                           | MathScore<br>0      | ReadingScore<br>0  | WritingScore<br>0           |                      |                     |                    |                        |
| Hay 1572 valores NA en la variable NrSiblings |                     |                    |                             |                      |                     |                    |                        |
| <pre>## {r} colSums(dataset == "") ##</pre>   |                     |                    |                             |                      |                     |                    |                        |
| Gender<br>0                                   | EthnicGroup<br>1840 | ParentEduc<br>1845 | ParentMaritalStatus<br>1190 | PracticeSport<br>631 | IsFirstChild<br>904 | NrSiblings<br>NA   | TransportMeans<br>3134 |
| WklyStudyHours<br>955                         | MathScore<br>0      | ReadingScore<br>0  | WritingScore<br>0           |                      |                     |                    |                        |

- c. **Transformación de los grupos étnicos:** La información contenida en los grupos étnicos viene codificada, la hemos transformado para permitir que sea más entendible.

```

{r}
equivalencias <- c("group A" = "Americana",
                  "group B" = "Negra",
                  "group C" = "Asiática",
                  "group D" = "Hispana",
                  "group E" = "Caucásica")

dataset$EthnicGroup <- equivalencias[dataset$EthnicGroup]

```

- d. **Dicotomización de dimensiones:** En concreto hemos decidido realizar la dicotomización sobre las variables “*IsFirstChild*”(Indica si es primer hijo), “*NrSiblings*”(indica si tiene hermanos o hermanas) y “*TransportMeans*” (indica el medio de transporte para ir al centro)

```

{r}
# Variable IsFirstChild
dataset$IsFirstChild <- ifelse(dataset$IsFirstChild == "yes", 0, 1)

# Variable NrSiblings
dataset$NrSiblings <- ifelse(dataset$NrSiblings == 0, 0, 1)

# Variable TransportMeans
dataset$TransportMeans <- ifelse(dataset$TransportMeans == "school_bus", 0, 1)

```

- e. **Nueva variable Nota:** Como último paso vamos a generar un nuevo valor como la media de las notas del alumno usando (“*MathScore*”, “*ReadingScore*” y “*WritingScore*”)

```

{r}
dataset$OverallScore <- round((dataset$MathScore + dataset$ReadingScore + dataset$WritingScore) / 3)

```

- f. **Transformación de horas de estudio:** Adicionalmente, se ha detectado un error de formato en los datos de *WklyStudyHours* y se ha transformado en un valor con sentido.

```

{r}
dataset$WklyStudyHours <- ifelse(dataset$WklyStudyHours == "5-oct", '5-10', dataset$WklyStudyHours)

```

- g. **Valores extremos:** Podemos ver que la distribución de las variables numéricas se parecen a la de una normal y que las puntuaciones bajas actúan como outliers. No vemos necesario aplicar ninguna medida respecto a los outliers, ya que parecen ser datos realistas dentro del rango de notas posible, por lo que eliminarlos podría suponer un análisis incorrecto de los datos.

## 4. Análisis de los datos

En este apartado realizaremos diferentes comprobaciones sobre los datos para detectar diferentes casuísticas.

**Variables seleccionadas para el analizar:** *Gender*, *EthnicGroup*, *ParentEduc*, *ParentMaritalStatus*, *PracticeSport*, *IsFirstChild*, *NrSiblings*, *TransportMeans*, *WklyStudyHours*, *MathScore*, *ReadingScore*, *WritingScore* y *OverallScore*

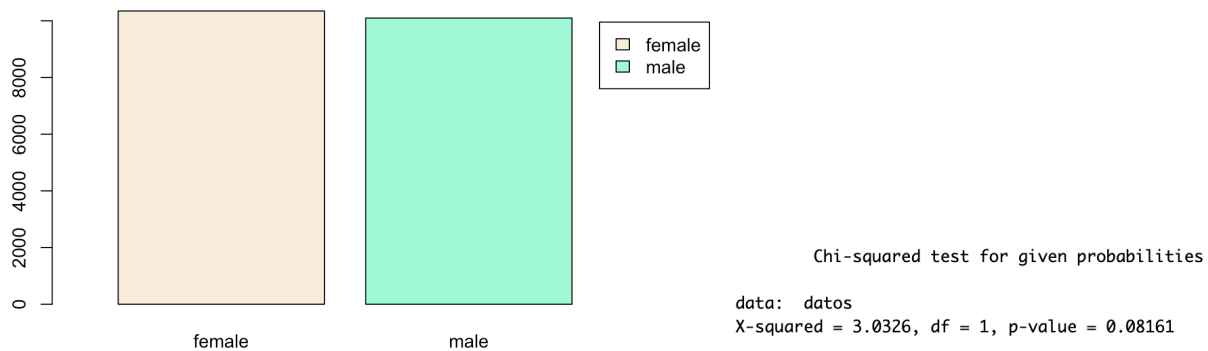
```
summary(dataset)
```

|                  |                  |                  |                     |                  |                |                |                |
|------------------|------------------|------------------|---------------------|------------------|----------------|----------------|----------------|
| Gender           | EthnicGroup      | ParentEduc       | ParentMaritalStatus | PracticeSport    | IsFirstChild   | NrSiblings     | TransportMeans |
| Length:20445     | Length:20445     | Length:20445     | Length:20445        | Length:20445     | Min. :0.0000   | Min. :0.0000   | Min. :0.0000   |
| Class :character | Class :character | Class :character | Class :character    | Class :character | 1st Qu.:0.0000 | 1st Qu.:1.0000 | 1st Qu.:0.0000 |
| Mode :character  | Mode :character  | Mode :character  | Mode :character     | Mode :character  | Median :0.0000 | Median :1.0000 | Median :0.0000 |
|                  |                  |                  |                     |                  | Mean :0.3548   | Mean :0.8973   | Mean :0.4124   |
|                  |                  |                  |                     |                  | 3rd Qu.:1.0000 | 3rd Qu.:1.0000 | 3rd Qu.:1.0000 |
|                  |                  |                  |                     |                  | Max. :1.0000   | Max. :1.0000   | Max. :1.0000   |

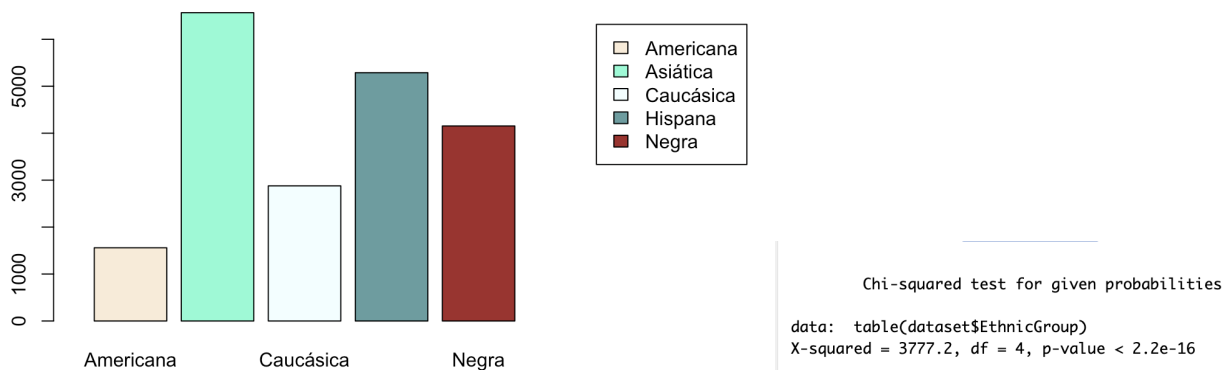
  

|                  |                |                |                |                |
|------------------|----------------|----------------|----------------|----------------|
| WklyStudyHours   | MathScore      | ReadingScore   | WritingScore   | OverallScore   |
| Length:20445     | Min. : 0.00    | Min. : 10.00   | Min. : 4.00    | Min. : 9.00    |
| Class :character | 1st Qu.: 56.00 | 1st Qu.: 60.00 | 1st Qu.: 58.00 | 1st Qu.: 58.00 |
| Mode :character  | Median : 67.00 | Median : 70.00 | Median : 69.00 | Median : 68.00 |
|                  | Mean : 66.65   | Mean : 69.55   | Mean : 68.59   | Mean : 68.26   |
|                  | 3rd Qu.: 78.00 | 3rd Qu.: 80.00 | 3rd Qu.: 79.00 | 3rd Qu.: 79.00 |
|                  | Max. :100.00   | Max. :100.00   | Max. :100.00   | Max. :100.00   |

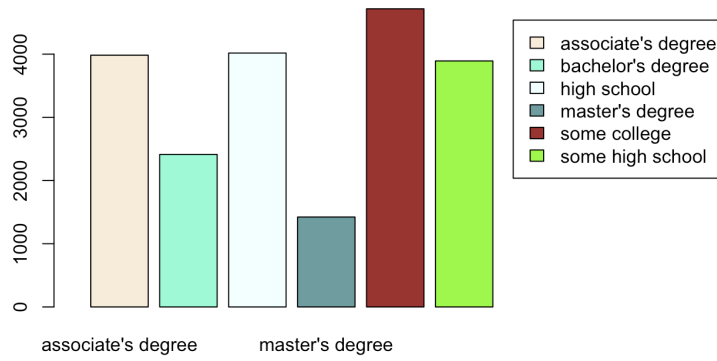
- a. **Gender:** Podemos comprobar que nuestra muestra tiene prácticamente el mismo número de mujeres que de hombres, siendo el de mujeres ligeramente superior. El Chi-cuadrado muestra que existe una asociación significativa entre los datos y las probabilidades dadas



- b. **EthnicGroup:** Respecto al grupo étnico, predominan los asiáticos e hispanos sobre el resto. Existe asociación altamente significativa



- c. **ParentEduc:** En cuanto a la educación de los padres, principalmente tienen formación profesional y el graduado escolar. Existe asociación altamente significativa

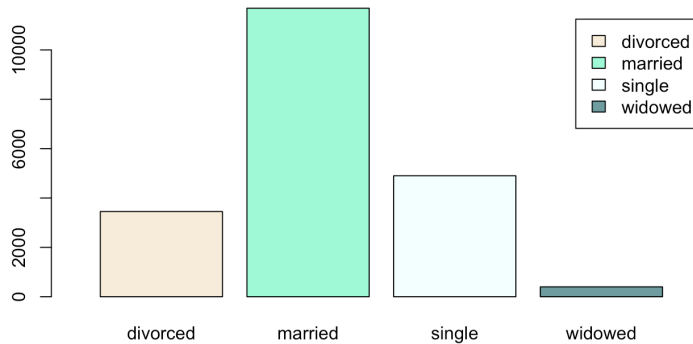


Chi-squared test for given probabilities

data: table(dataset\$ParentEduc)

X-squared = 2224.4, df = 5, p-value < 2.2e-16

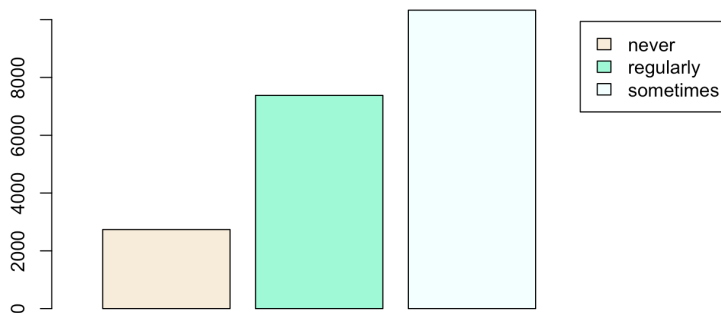
- d. **ParentMaritalStatus:** Destacan sobre todo los padres que están casados. La asociación también es altamente significativa.



Chi-squared test for given probabilities

data: table(dataset\$ParentMaritalStatus)  
X-squared = 13359, df = 3, p-value < 2.2e-16

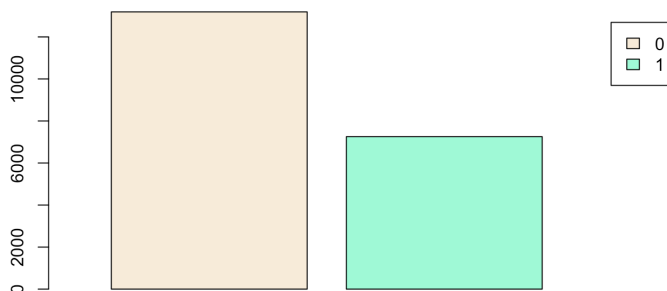
- e. **PracticeSport:** La mayoría de los individuos de la muestra practican algún deporte de vez en cuando. Volvemos a tener fuerte asociación



Chi-squared test for given probabilities

data: datos  
X-squared = 4297.9, df = 2, p-value < 2.2e-16

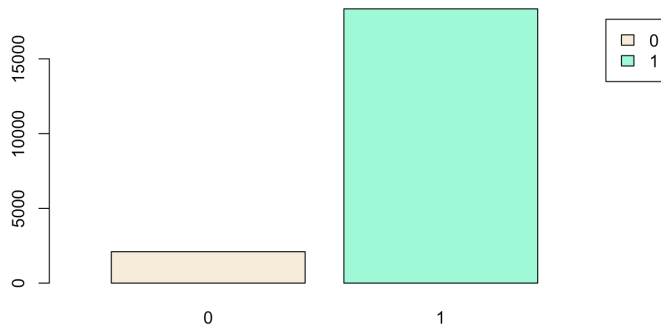
- f. **IsFirstChild:** La mayoría de los individuos son el primer hijo de la familia. Asociación significativa



Chi-squared test for given probabilities

data: datos  
X-squared = 1724, df = 1, p-value < 2.2e-16

- g. NrSiblings:** La mayoría de la muestra tiene hermanos y hermanas. Asociación significativa

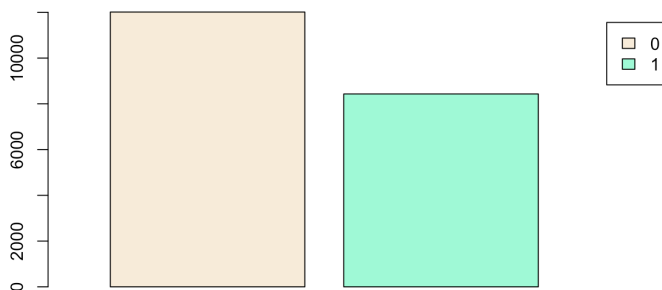


Chi-squared test for given probabilities

data: datos

X-squared = 12908, df = 1, p-value < 2.2e-16

- h. TransportMeans:** La mayoría se desplaza al centro educativo mediante transporte escolar. Fuerte asociación



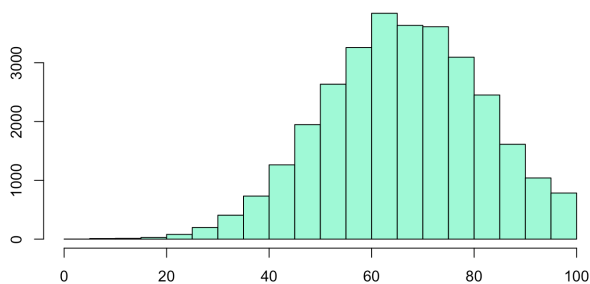
Chi-squared test for given probabilities

data: datos

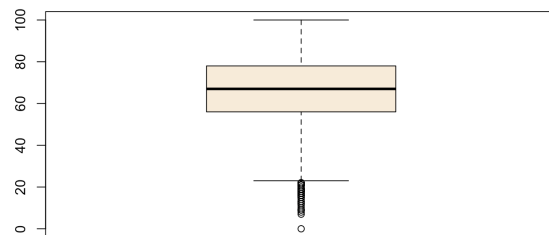
X-squared = 627.22, df = 1, p-value < 2.2e-16

- i. MathScore:** La mediana de puntuaciones en matemáticas es de 70, tenemos outliers por debajo de los 20 puntos

Distribución MathScore

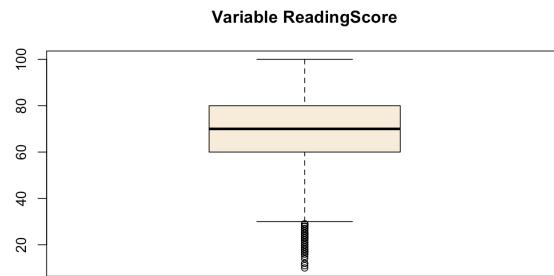
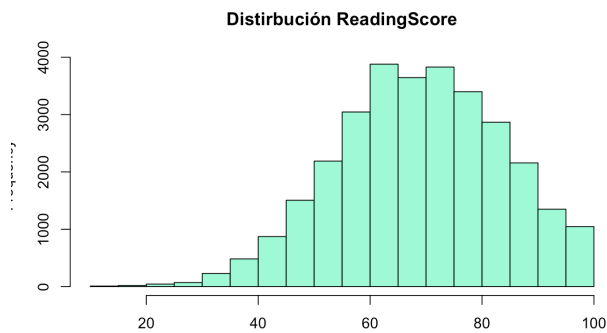


Variable MathScore

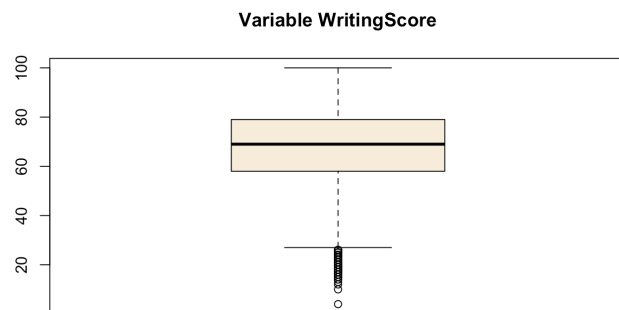
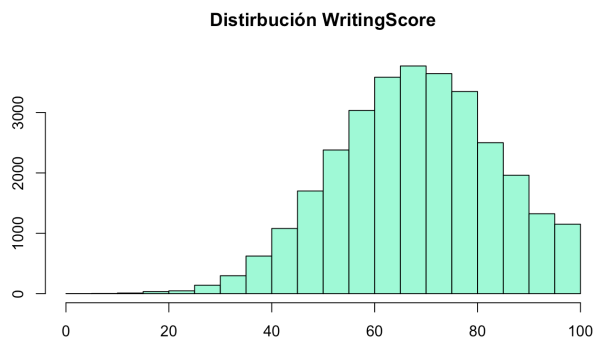


- j. ReadingScore:** La mediana es similar a la de matemáticas para las puntuaciones en lectura en torno a 70. En este caso se consideran outliers los valores inferiores a 30.

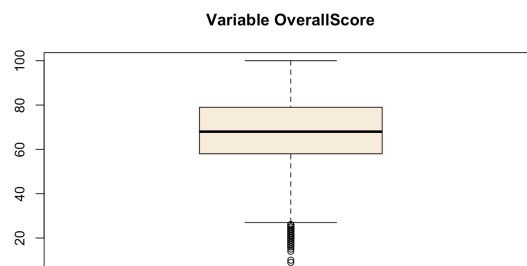
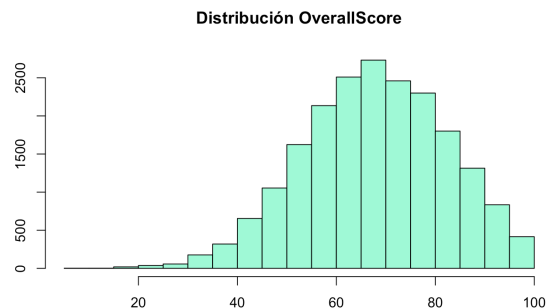




**k. WritingScore:** Volvemos a tener una mediana en torno a los 70 y outliers por debajo de los 30

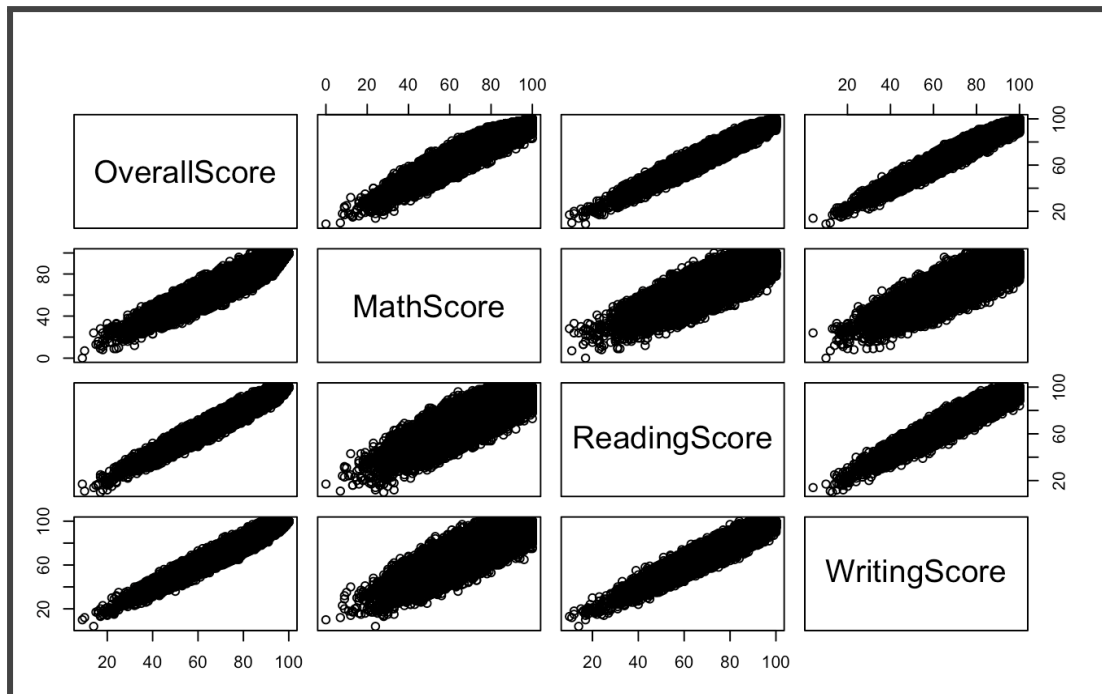


**l. OverallScore:** En vista a los anteriores resultados las puntuaciones medias siguen la misma tendencias, con una mediana en torno a los 70 y outliers por debajo de los 30



## Regresión lineal sobre los datos:

Modelo de regresión lineal simple:



## Modelo de regresión logística:

```
call:
glm(formula = OverallScore_sobre ~ Gender + EthnicGroup + ParentEduc +
  Practicesport, data = dataset)
```

Coefficients:

|                             | Estimate  | Std. Error | t value | Pr(> t ) |     |
|-----------------------------|-----------|------------|---------|----------|-----|
| (Intercept)                 | 0.120004  | 0.012164   | 9.866   | < 2e-16  | *** |
| Gendermale                  | -0.065085 | 0.004878   | -13.342 | < 2e-16  | *** |
| EthnicGroupAsiática         | 0.019044  | 0.009832   | 1.937   | 0.052758 | .   |
| EthnicGroupCaucásica        | 0.148733  | 0.010969   | 13.559  | < 2e-16  | *** |
| EthnicGroupHispana          | 0.071954  | 0.010058   | 7.154   | 8.73e-13 | *** |
| EthnicGroupNegra            | 0.006404  | 0.010362   | 0.618   | 0.536549 |     |
| ParentEducbachelor's degree | 0.052976  | 0.008969   | 5.907   | 3.55e-09 | *** |
| ParentEduchigh school       | -0.070504 | 0.007801   | -9.038  | < 2e-16  | *** |
| ParentEducmaster's degree   | 0.097900  | 0.010754   | 9.104   | < 2e-16  | *** |
| ParentEducsome college      | -0.038796 | 0.007509   | -5.166  | 2.41e-07 | *** |
| ParentEducsome high school  | -0.095161 | 0.007862   | -12.105 | < 2e-16  | *** |
| Practicesportregularly      | 0.055877  | 0.007803   | 7.161   | 8.30e-13 | *** |
| Practicesportsometimes      | 0.027639  | 0.007495   | 3.687   | 0.000227 | *** |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.11441)

Null deviance: 2336.2 on 19242 degrees of freedom  
Residual deviance: 2200.1 on 19230 degrees of freedom  
AIC: 12906

Number of Fisher scoring iterations: 2

## 5. Resolución del problema

La creación de un modelo logístico nos ha permitido entender mejor las variables que influyen en un mayor rendimiento académico. Las conclusiones obtenidas son las siguientes:

- Ser hombre está asociado con una menor probabilidad de tener una puntuación alta en "OverallScore" en comparación con ser mujer.
- El origen étnico caucásico está relacionado con un mayor rendimiento académico respecto a los otros orígenes étnicos.
- Respeco a la educación de las padres, si estos tienen formación de máster está relacionado con un mayor rendimiento académico, sin embargo si los padres tienen simplemente el graduado escolar, está relacionado con un menor rendimiento
- Practicar deporte regularmente está asociado con una mayor probabilidad de tener una puntuación alta en OverallScore.

