

Práctica II

Tipología

y

Ciclo de Vida de los datos



GREGORIO ANDRÉS GARCÍA MENÉNDEZ
MANUEL GÓMEZ MONTERO

8 de junio de 2019

Índice

1. Introducción	3
1.1. Objetivos	3
1.2. Descripción del Dataset	3
2. Limpieza de datos	5
2.1. Unión de conjuntos	5
2.2. Tratamiento de NAs	6
2.3. Análisis de 0s	8
2.4. Selección de variables/Reducción de dimensionalidad	8
2.5. Tipos de Variables	9
2.6. Análisis de Outliers	10

3. Análisis Estadístico	12
3.1. Análisis gráfico inicial	12
3.2. Estadística Inferencial	16
3.2.1. Estudiantes de distinto sexo	16
3.2.2. Estudiantes con diferente situación de convivencia de los padres	17
3.2.3. Estudiantes que aprueban ($G \geq 10$) y estudiantes que suspenden ($G < 10$)	19
3.3. Modelo de regresión lineal	20
4. Estudio de correlación	21
5. Conclusiones	23

1. Introducción

1.1. Objetivos

Los objetivos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

1.2. Descripción del Dataset

El dataset escogido ha sido obtenido a través de una encuesta de estudiantes de secundaria que asisten a cursos de matemáticas y portugués. Este conjunto de datos contiene bastante información de interés sobre los estudiantes. Aunque inicialmente el objetivo principal del dataset era ver el consumo de alcohol en estudiantes según variables sociodemográficas, dicho dataset contiene suficiente información para otro tipo de estudios, centrados por ejemplo en la calificación de los alumnos.

Los datos han sido obtenidos de *Kaggle* y se encuentran en dos ficheros *CSV* independientes, uno para la asignatura de matemáticas y otro para la de portugués. Cada uno de estos ficheros contiene las siguientes variables:

1. *school*: Define la escuela del estudiante. Puede ser “GP” (Gabriel Pereira) o “MS” (Mousinho da Silveira).
2. *sex*: Indica el sexo del estudiante.
3. *age*: Se refiere a la edad del estudiante.
4. *address*: Indica si el estudiante vive en zona urbana o rural.
5. *famsize*: Define si la familia se compone de menos de tres miembros o de tres o más.
6. *Pstatus*: Se refiere al estado de los padres, si viven juntos o no.
7. *Medu*: Nivel de educación de la madre (0: Ninguna, 1: Primaria (hasta 4º), 2: Primaria (Desde 5º), 3: Secundaria o 4: Educación superior).
8. *Fedu*: Nivel de educación del padre (0: Ninguna, 1: Primaria (hasta 4º), 2: Primaria (Desde 5º), 3: Secundaria o 4: Educación superior).
9. *Mjob*: Trabajo de la madre.
10. *Fjob*: Trabajo del padre.
11. *reason*: Razón por la que se escoge esta escuela.
12. *guardian*: Tutor del alumno.
13. *traveltime*: Tiempo de camino a la escuela (1: ¡15 min., 2: 15 to 30 min., 3: 30 min. a 1 hora, 4: ¡1 hora).
14. *studytime*: Tiempo de estudio semanal (1: ¡2 horas, 2: 2 a 5 horas, 3: 5 a 10 horas o 4: ¡10 horas).
15. *failures* - Número de fracasos en clases pasadas (4 para 4 o más)
16. *schoolsup* - Define si el estudiante recibe o no clases particulares.
17. *famsup* - Indica si el estudiante recibe apoyo educativo familiar.
18. *paid*: Clases extra pagadas dentro de la asignatura del curso.
19. *activities*: Actividades extraescolares.
20. *nursery*: Indica si asistió a la guardería.
21. *higher*: Define si el estudiante tiene intención de realizar estudios superiores.

- 22. *internet*: Indica si el estudiante posee internet en casa.
- 23. *romantic*: Define si el estudiante tiene una relación amorosa.
- 24. *famrel*: Calidad de las relaciones familiares (De 1 a 5 de peor a mejor).
- 25. *freetime*: Indica la cantidad de tiempo libre del estudiante (De 1 a 5 de menos a más).
- 26. *goout*: Mide la frecuencia con la que el estudiante sale con amigos (De 1 a 5 de menos a más).
- 27. *Dalc* -Consumo de alcohol entre semana (De 1 a 5 de menos a más).
- 28. *Walc* - Consumo de alcohol los fines de semana (De 1 a 5 de menos a más).
- 29. *health* - Mide el estado de salud del estudiante (De 1 a 5 de peor a mejor).
- 30. *absences*: Número de ausencias a clase.
- 31. *G1*: Calificación del primer periodo.
- 32. *G2*: Calificación del segundo periodo.
- 33. *G3*: Calificación final.

2. Limpieza de datos

2.1. Unión de conjuntos

El primer paso consiste en unir ambos conjuntos de datos de cada clase para tener un único dataset que será el que utilizaremos en el resto del documento.

Según se indica en Kaggle existen 382 estudiantes que asisten a ambos cursos. Como cada alumno en los dos ficheros no viene identificado con un *id* único, explican que estos estudiantes pueden ser identificados por el mismo valor de las siguientes características en ambos ficheros: colegio, sexo, edad, dirección, tamaño de familia, trabajo y nivel de educación de los padres, razón por la que ha elegido el colegio, si han asistido a la guardería y si poseen internet en casa.

Sin embargo creemos que además de estas características hay otras que no pueden variar en un mismo estudiante en ambos ficheros como son: el tiempo de camino a la escuela, número de fracasos en clases pasadas, si realiza actividades extraescolares, si tiene intención de realizar estudios superiores, si tiene una relación, la calidad de las relaciones, la frecuencia de salidas con amigos, la cantidad consumida de alcohol y el estado de salud. Con esta separación vemos que existen 320 estudiantes que asisten a ambas clases.

Comparando los valores de ambos cursos del resto del variables vemos que también coinciden en ambos cursos el tutor del alumno, si recibe soporte familiar y el tiempo de estudio.

school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason
GP:485	F:417	Min. :15.00	R:218	GT3:508	A: 90	Min. :0.000	Min. :0.000	at home :150	at home : 48	course :312
MS:239	M:307	1st Qu.:16.00	U:506	LE3:216	T:634	1st Qu.:2.000	1st Qu.:1.000	health : 52	health : 26	home :171
		Median :17.00				Median :2.000	Median :2.000	other :283	other :407	other : 78
		Mean :16.81				Mean :2.485	Mean :2.285	services:164	services:205	reputation:163
		3rd Qu.:18.00				3rd Qu.:4.000	3rd Qu.:3.000	teacher : 75	teacher : 38	
		Max. :22.00				Max. :4.000	Max. :4.000			

nursery	internet	traveltime	failures	schoolsup	activities	higher	romantic	famrel	freetime	goout
no :150	no :169	Min. :1.000	Min. :0.0000	no :648	no :378	no : 82	no :452	Min. :1.000	Min. :1.0	Min. :1.000
yes:574	yes:555	1st Qu.:1.000	1st Qu.:0.0000	yes: 76	yes:346	yes:642	yes:272	1st Qu.:4.000	1st Qu.:3.0	1st Qu.:2.000
		Median :1.000	Median :0.0000					Median :4.000	Median :3.0	Median :3.000
		Mean :1.565	Mean :0.3453					Mean :3.913	Mean :3.2	Mean :3.195
		3rd Qu.:2.000	3rd Qu.:0.0000					3rd Qu.:5.000	3rd Qu.:4.0	3rd Qu.:4.000
		Max. :4.000	Max. :3.0000					Max. :5.000	Max. :5.0	Max. :5.000

Dalc	Walc	health	guardian	famsup	studytime	paid.mat	absences.mat	G1.mat
Min. :1.000	Min. :1.000	Min. :1.000	father:169	no :287	Min. :1.00	no :214	Min. : 0.000	Min. : 3.00
1st Qu.:1.000	1st Qu.:1.000	1st Qu.:2.000	mother:491	yes:437	1st Qu.:1.00	yes :181	1st Qu.: 0.000	1st Qu.: 8.00
Median :1.000	Median :2.000	Median :4.000	other : 64		Median :2.00	NA's:329	Median : 4.000	Median :11.00
Mean :1.519	Mean :2.311	Mean :3.552			Mean :1.92		Mean : 5.709	Mean :10.91
3rd Qu.:2.000	3rd Qu.:3.000	3rd Qu.:5.000			3rd Qu.:2.00		3rd Qu.: 8.000	3rd Qu.:13.00
Max. :5.000	Max. :5.000	Max. :5.000			Max. :4.00		Max. :75.000	Max. :19.00

G2.mat	G3.mat	paid.por	absences.por	G1.por	G2.por	G3.por
Min. : 0.00	Min. : 0.00	no :610	Min. : 0.000	Min. : 0.0	Min. : 0.00	Min. : 0.00
1st Qu.: 9.00	1st Qu.: 8.00	yes : 39	1st Qu.: 0.000	1st Qu.:10.0	1st Qu.:10.00	1st Qu.:10.00
Median :11.00	Median :11.00	NA's: 75	Median : 2.000	Median :11.0	Median :11.00	Median :12.00
Mean :10.71	Mean :10.42		Mean : 3.659	Mean :11.4	Mean :11.57	Mean :11.91
3rd Qu.:13.00	3rd Qu.:14.00		3rd Qu.: 6.000	3rd Qu.:13.0	3rd Qu.:13.00	3rd Qu.:14.00
Max. :19.00	Max. :20.00		Max. :32.000	Max. :19.0	Max. :19.00	Max. :19.00
NA's :329	NA's :329		NA's :75	NA's :75	NA's :75	NA's :75

Figura 1: Summary dataset conjunto

El dataset resultado de la unión posee un total de 724 alumnos 75 de los cuales solo asisten a matemáticas, 329 solo a portugués y 320 a ambos cursos.

2.2. Tratamiento de NAs

En primer lugar vamos a encargarnos de los valores perdidos que se dan unicamente en los casos de que los alumnos no hayan asistido a alguno de los cursos, una opción sería eliminar los datos de cualquier alumno que no posea datos en algunos de los cursos, pero implicaría quedarnos solo con un 44% de los datos.

Por lo tanto la opción escogida para los valores perdidos es la siguiente:

```

students.nonas <- students.merge[, c("school","sex","age","address","famsize","Pstatus","Medu","Fedu",
", "Mjob", "Fjob", "reason", "nursery", "internet", "traveltime", "failures", "schoolsup", "activities",
", "higher", "romantic", "famrel", "freetime", "goout", "Dalc", "Walc", "health", "guardian", "famsup",
", "studytime", "G1.por", "G1.mat", "G2.por", "G2.mat", "G3.por", "G3.mat", "absences.mat", "absences.
por")]
students.nonas$paid <- "no"
students.nonas$paid[NVL(students.merge$paid.mat == 'yes' | students.merge$paid.por == 'yes', FALSE)]
<- "yes"
students.nonas$paid <- as.factor(students.nonas$paid)

summary(students.nonas)

```

Figura 2: Unificación variable paid

- *paid*: Unificaremos las variables de ambos cursos tomando un valor “yes” cuando el alumno de clases extras pagadas de alguno de los cursos.
- *calificaciones* y *absences*: Utilizaremos el método *missForest* que utiliza árboles de decisión para imputar los valores perdidos en variables tanto numéricas como categóricas.

```

mf <- missForest(students.nonas, maxiter = 10, ntree = 100, variablewise = FALSE,
decreasing = FALSE, verbose = FALSE,
mtry = floor(sqrt(ncol(students.nonas))), replace = TRUE,
classwt = NULL, cutoff = NULL, strata = NULL,
sampsize = NULL, nodesize = NULL, maxnodes = NULL,
xtrue = NA, parallelize = c('no', 'variables', 'forests'))

```

Figura 3: Imputación valores perdidos con missForest

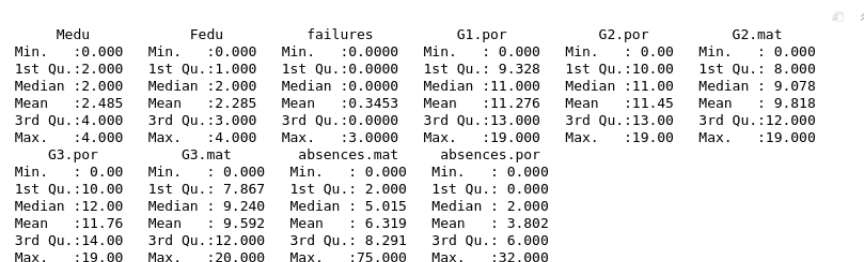
school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob
GP:485	F:417	Min. :15.00	R:218	GT3:508	A: 90	Min. :0.000	Min. :0.000	at_home :150	at_home : 48
MS:239	M:307	1st Qu.:16.00	U:506	LE3:216	T:634	1st Qu.:2.000	1st Qu.:1.000	health : 52	health : 26
		Median :17.00				Median :2.000	Median :2.000	other :283	other :407
		Mean :16.81				Mean :2.485	Mean :2.285	services:164	services:205
		3rd Qu.:18.00				3rd Qu.:4.000	3rd Qu.:3.000	teacher : 75	teacher : 38
		Max. :22.00				Max. :4.000	Max. :4.000		
reason	nursery	internet	traveltime	failures	schoolsup	activities	higher	romantic	famrel
course :312	no :150	no :169	Min. :1.000	Min. :0.0000	no :648	no :378	no : 82	no :452	Min. :1.000
home :171	yes:574	yes:555	1st Qu.:1.000	1st Qu.:0.0000	yes: 76	yes:346	yes:642	yes:272	1st Qu.:4.000
other : 78			Median :1.000	Median :0.0000					Median :4.000
reputation:163			Mean :1.565	Mean :0.3453					Mean :3.913
			3rd Qu.:2.000	3rd Qu.:0.0000					3rd Qu.:5.000
			Max. :4.000	Max. :3.0000					Max. :5.000
freetime	goout	Dalc	Walc	health	guardian	famsup	studytime	G1.por	
Min. :1.0	Min. :1.000	Min. :1.000	Min. :1.000	Min. :1.000	father:169	no :287	Min. :1.00	Min. : 0.000	
1st Qu.:3.0	1st Qu.:2.000	1st Qu.:1.000	1st Qu.:1.000	1st Qu.:2.000	mother:491	yes:437	1st Qu.:1.00	1st Qu.: 9.328	
Median :3.0	Median :3.000	Median :1.000	Median :2.000	Median :4.000	other : 64		Median :2.00	Median :11.000	
Mean :3.2	Mean :3.195	Mean :1.519	Mean :2.311	Mean :3.552			Mean :1.92	Mean :11.276	
3rd Qu.:4.0	3rd Qu.:4.000	3rd Qu.:2.000	3rd Qu.:3.000	3rd Qu.:5.000			3rd Qu.:2.00	3rd Qu.:13.000	
Max. :5.0	Max. :5.000	Max. :5.000	Max. :5.000	Max. :5.000			Max. :4.00	Max. :19.000	
G1.mat	G2.por	G2.mat	G3.por	G3.mat	absences.mat	absences.por	paid		
Min. : 3.000	Min. : 0.00	Min. : 0.000	Min. : 0.00	Min. : 0.000	Min. : 0.000	Min. : 0.000	no :515		
1st Qu.: 8.000	1st Qu.:10.00	1st Qu.: 8.000	1st Qu.:10.00	1st Qu.: 7.867	1st Qu.: 2.000	1st Qu.: 0.000	yes:209		
Median : 9.207	Median :11.00	Median : 9.078	Median :12.00	Median : 9.240	Median : 5.015	Median : 2.000			
Mean :10.048	Mean :11.45	Mean : 9.818	Mean :11.76	Mean : 9.592	Mean : 6.319	Mean : 3.802			
3rd Qu.:12.000	3rd Qu.:13.00	3rd Qu.:12.000	3rd Qu.:14.00	3rd Qu.:12.000	3rd Qu.: 8.291	3rd Qu.: 6.000			
Max. :19.000	Max. :19.00	Max. :19.000	Max. :19.00	Max. :20.000	Max. :75.000	Max. :32.000			

Figura 4: Summary dataset sin NAs

En el notebook se pueden ver los detalles de estos procesos en la figura 4 podemos ver la salida del comando *summary* para este nuevo dataset.

2.3. Análisis de 0s

Vamos a analizar uno por uno los casos en los que existen valores 0 y a definir si son valores posibles de la variable o por el contrario se trata de valores vacíos indicados como 0. Las variables que contienen valores 0 son:



```

      Medu      Fedu      failures      G1.por      G2.por      G2.mat
Min.   :0.000  Min.   :0.000  Min.   :0.0000  Min.   : 0.000  Min.   : 0.00  Min.   : 0.000
1st Qu.:2.000  1st Qu.:1.000  1st Qu.:0.0000  1st Qu.: 9.328  1st Qu.:10.00  1st Qu.: 8.000
Median :2.000  Median :2.000  Median :0.0000  Median :11.000  Median :11.00  Median : 9.078
Mean   :2.485  Mean   :2.285  Mean   :0.3453  Mean   :11.276  Mean   :11.45  Mean   : 9.818
3rd Qu.:4.000  3rd Qu.:3.000  3rd Qu.:0.0000  3rd Qu.:13.000  3rd Qu.:13.00  3rd Qu.:12.000
Max.   :4.000  Max.   :4.000  Max.   :3.0000  Max.   :19.000  Max.   :19.00  Max.   :19.000

      G3.por      G3.mat      absences.mat      absences.por
Min.   : 0.00  Min.   : 0.000  Min.   : 0.000  Min.   : 0.000
1st Qu.:10.00  1st Qu.: 7.867  1st Qu.: 2.000  1st Qu.: 0.000
Median :12.00  Median : 9.240  Median : 5.015  Median : 2.000
Mean   :11.76  Mean   : 9.592  Mean   : 6.319  Mean   : 3.802
3rd Qu.:14.00  3rd Qu.:12.000  3rd Qu.: 8.291  3rd Qu.: 6.000
Max.   :19.00  Max.   :20.000  Max.   :175.000  Max.   :32.000

```

Figura 5: Summary atributos con 0s

Para el caso de la educación del padre y la madre ya definimos que el valor 0 significa que no poseen ningún tipo de educación.

Para el resto de casos vemos que el valor 0 también es posible ya que:

- Es posible no haber suspendido ninguna asignatura, de hecho más del 75 % de alumnos así lo han hecho.
- En las calificaciones es posible sacar un 0.
- En el caso de las ausencias también existen alumnos que no han faltado a ninguna clase. En el caso
- de las clases de portugués más del 25 %.

2.4. Selección de variables/Reducción de dimensionalidad

Observando el dataset vemos que, aún tras la unión de los datasets, tenemos una gran cantidad de dimensiones, en concreto 37, lo que puede dificultar nuestro análisis. En esta primera fase vamos a hacer una primera aproximación para reducir la dimensionalidad. Posteriormente en los primeros apartados del análisis estadístico reduciremos aún más esta dimensionalidad.

En primer lugar existen variables que, a priori, tienen poca significancia para nuestro análisis como puede ser la escuela a la que asiste el alumno, por lo que podemos eliminar esta variable.

Por otro lado podemos agrupar las calificaciones en una única variable que nos indique la calificación media del alumno.

```
students.red$G <- rowMeans(students.nonas[c('G1.mat', 'G1.por', 'G2.mat', 'G2.por', 'G3.mat', 'G3.por')])
```

Figura 6: Unificación calificaciones

También vamos a realizar el mismo paso para las audiencias, pero observando los datos parece que las ausencias son mayores en matemáticas. Como lo que nos interesa es saber los alumnos que han faltado más o menos veces y queremos evitar que tenga un mayor peso los alumnos de matemáticas, normalizaremos estos valores antes de realizar la media.

```
normalized<-function(y) {  
  x<-y[!is.na(y)]  
  x<-(x - min(x)) / (max(x) - min(x))  
  y[!is.na(y)]<-x  
  return(y)  
}  
  
students.red$absences <- rowMeans(sapply(students.nonas[c('absences.mat', 'absences.por')], normalized), na.rm=TRUE)  
  
summary(students.red)
```

Figura 7: Unificación ausencias

Con estos detalles que pueden observarse en el notebook vemos que hemos conseguido reducir a 30 la dimensionalidad con las variables que aparecen en la figura 8.

2.5. Tipos de Variables

En los tipos mostrados hay algunos que no es correcto el tipo. En la educación tanto de la madre como del padre consideramos que aunque se use un número para la representación debería ser un factor. Lo mismo ocurre para el tiempo de viaje o el tiempo de estudio. Podemos encontrar más detalles en el notebook.

school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob
GP:485	F:417	Min. :15.00	R:218	GT3:508	A: 90	Min. :0.000	Min. :0.000	at_home :150	at_home : 48
MS:239	M:307	1st Qu.:16.00	U:506	LE3:216	T:634	1st Qu.:2.000	1st Qu.:1.000	health : 52	health : 26
		Median :17.00				Median :2.000	Median :2.000	other :283	other :407
		Mean :16.81				Mean :2.485	Mean :2.285	services:164	services:205
		3rd Qu.:18.00				3rd Qu.:4.000	3rd Qu.:3.000	teacher : 75	teacher : 38
		Max. :22.00				Max. :4.000	Max. :4.000		
reason	nursery	internet	traveltime	failures	schoolsup	activities	higher	romantic	famrel
course :312	no :150	no :169	Min. :1.000	Min. :0.0000	no :648	no :378	no : 82	no :452	Min. :1.000
home :171	yes:574	yes:555	1st Qu.:1.000	1st Qu.:0.0000	yes: 76	yes:346	yes:642	yes:272	1st Qu.:4.000
other : 78			Median :1.000	Median :0.0000					Median :4.000
reputation:163			Mean :1.565	Mean :0.3453					Mean :3.913
			3rd Qu.:2.000	3rd Qu.:0.0000					3rd Qu.:5.000
			Max. :4.000	Max. :3.0000					Max. :5.000
freetime	goout	Dalc	Walc	health	guardian	famsup	studytime	G1.por	
Min. :1.0	Min. :1.000	Min. :1.000	Min. :1.000	Min. :1.000	father:169	no :287	Min. :1.00	Min. : 0.000	
1st Qu.:3.0	1st Qu.:2.000	1st Qu.:1.000	1st Qu.:1.000	1st Qu.:2.000	mother:491	yes:437	1st Qu.:1.00	1st Qu. : 9.328	
Median :3.0	Median :3.000	Median :1.000	Median :2.000	Median :4.000	other : 64		Median :2.00	Median :11.000	
Mean :3.2	Mean :3.195	Mean :1.519	Mean :2.311	Mean :3.552			Mean :1.92	Mean :11.276	
3rd Qu.:4.0	3rd Qu.:4.000	3rd Qu.:2.000	3rd Qu.:3.000	3rd Qu.:5.000			3rd Qu.:2.00	3rd Qu.:13.000	
Max. :5.0	Max. :5.000	Max. :5.000	Max. :5.000	Max. :5.000			Max. :4.00	Max. :19.000	
G1.mat	G2.por	G2.mat	G3.por	G3.mat	absences.mat	absences.por	paid		
Min. : 3.000	Min. : 0.00	Min. : 0.000	Min. : 0.00	Min. : 0.000	Min. : 0.000	Min. : 0.000	no :515		
1st Qu.: 8.000	1st Qu.:10.00	1st Qu.: 8.000	1st Qu.:10.00	1st Qu.: 7.867	1st Qu.: 2.000	1st Qu.: 0.000	yes:209		
Median : 9.207	Median :11.00	Median : 9.078	Median :12.00	Median : 9.240	Median : 5.015	Median : 2.000			
Mean :10.048	Mean :11.45	Mean : 9.818	Mean :11.76	Mean : 9.592	Mean : 6.319	Mean : 3.802			
3rd Qu.:12.000	3rd Qu.:13.00	3rd Qu.:12.000	3rd Qu.:14.00	3rd Qu.:12.000	3rd Qu.: 8.291	3rd Qu.: 6.000			
Max. :19.000	Max. :19.00	Max. :19.000	Max. :19.00	Max. :20.000	Max. :75.000	Max. :32.000			

Figura 8: Summary dataset reducido

La explicación de considerar el tiempo de estudio y el tiempo de viaje es que, aunque pueden parecer numéricas, en realidad son factores. En la descripción del dataset aparece explicado:

studytime: Weekly study time: (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours) traveltime: Home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

Por lo tanto, al ser categorías, las tratamos como factores y no como numéricas.

2.6. Análisis de Outliers

Para las variables numéricas realizamos un estudio de outliers. En nuestro caso, las variables a estudiar son “age” y “G”. El resto de variables numéricas en realidad no lo son, puesto que son categóricas. Por ejemplo: “studytime” va de 1 a 4, y no hace referencia a las horas que el alumno pasa estudiando, sino que son categorías equivalentes a por ejemplo: Nada, Poco, Normal, Mucho. Podría haber valores erróneos debido a la transcripción de los datos o similar, pero dichos errores ya los hubiéramos detectado en la creación del data set, ya que gracias a la función “summary” vemos los valores mínimos y máximos y para estas variables categóricas numéricas no hay valores erróneos (mínimo y máximo corresponden a las categorías mínimas y máximas).

Procedemos al estudio de age y G:

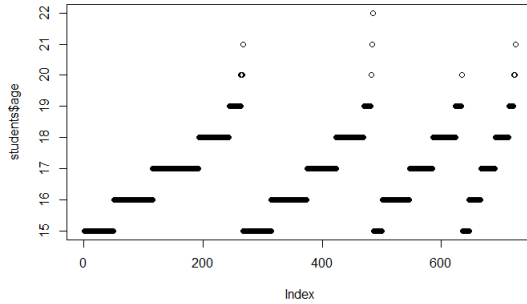


Figura 9: Distribución de la variable Age

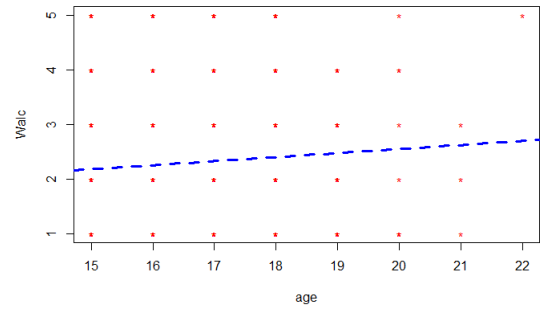


Figura 10: Walc en función de Age con outliers. Modelo lineal en azul.

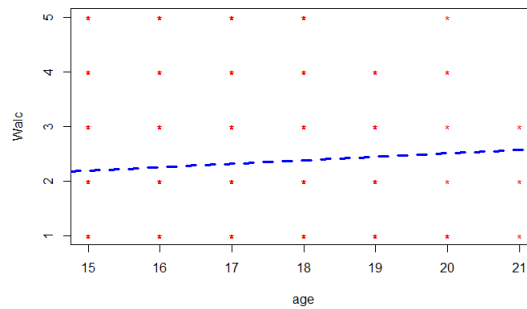


Figura 11: Walc en función de Age sin outliers. Modelo lineal en azul.

Como podemos ver, para el campo edad los alumnos van de los 15 a los 22, siendo las franjas más pobladas los 15, 16 y 17 años. Al haber valores intermedios que en número van disminuyendo gradualmente desde los 18 hasta los 22, no consideramos ningún valor extremo (como el único estudiantes de 22 años) como valor erróneo. Además, al ser tan pocos estudiantes, no los descartamos en nuestros estudios ya que pueden aportar información valiosa, y como podemos ver en la comparativa del conjunto con outliers y sin outliers, no cambia de forma crítica.

En el caso de la media de la nota G es más claro todavía. Los datos no demuestran valores extremos que se puedan deber a errores, y aquellos más alejados de la mayor concentración de estudiantes son valores que aportan información para los estudios que vamos a realizar en cuanto al consumo de alcohol y de desempeño estudiantil.

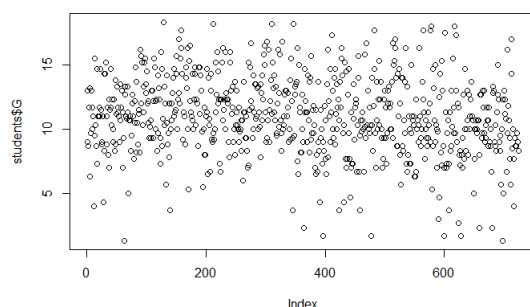


Figura 12: Distribución de la variable G

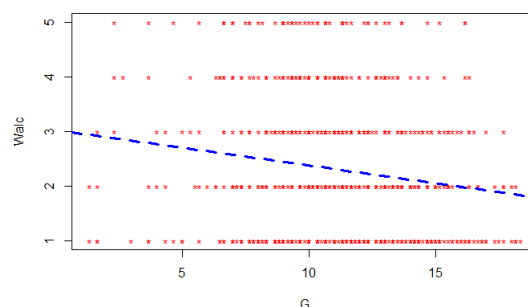


Figura 13: Walc en función de G. Modelo lineal en azul.

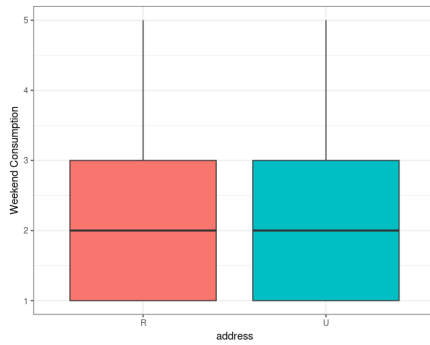
3. Análisis Estadístico

3.1. Análisis gráfico inicial

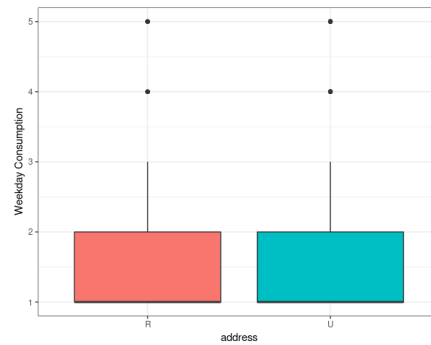
Observando las gráficas vemos que hay variables que parece que sí tienen influencia a simple vista en el consumo de alcohol tanto a diario como los fines de semana, como pueden ser el sexo y el estado de los padres. Sin embargo vemos otros que, a priori, no parece que tengan influencia así que en un primer momento vamos a dejar fuera del análisis la dirección, si el alumno tiene internet, si tiene una relación, el tutor, si realiza actividades extraescolares o si recibe clases de pago.

En este apartado vamos a mostrar únicamente las gráficas de las variables que hemos descartado en el estudio inicial, para ver que, efectivamente, a priori no tienen influencia en el consumo de alcohol. Las gráficas completas se muestran en el *notebook*.

En la figura 20 podemos ver el resumen del dataset. Tras la fase de limpieza y este pequeño análisis hemos pasado de tener dos datasets con 33 dimensiones a **un único dataset con 21 dimensiones** pasando por un dataset conjunto de 41 dimensiones.

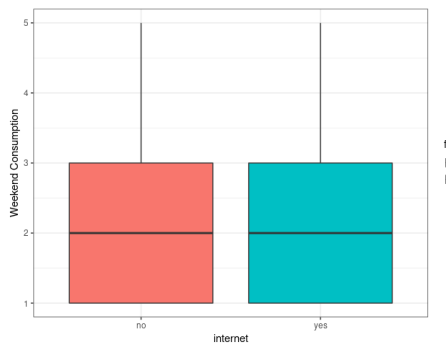


(a) Fin de semana

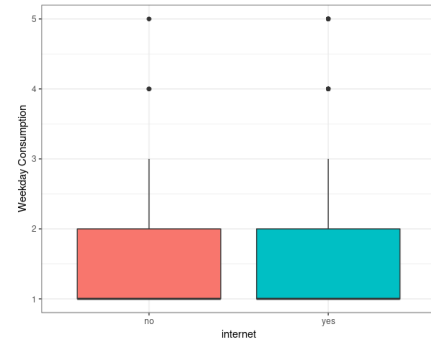


(b) Diario

Figura 14: Boxplot dirección - consumo alcohol

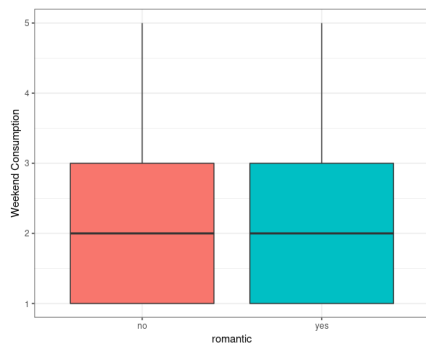


(a) Fin de semana

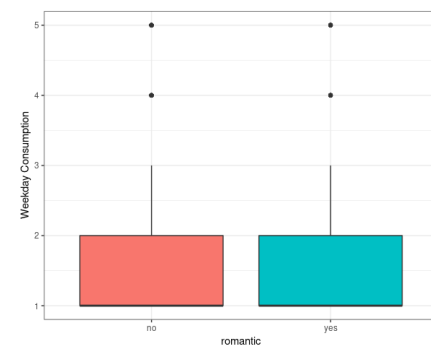


(b) Diario

Figura 15: Boxplot Internet - consumo alcohol

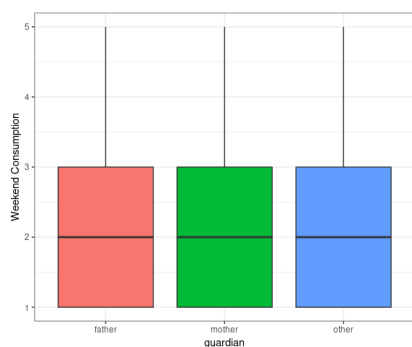


(a) Fin de semana

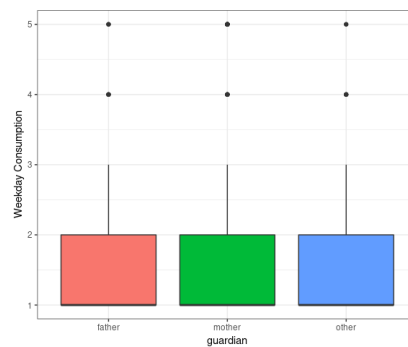


(b) Diario

Figura 16: Boxplot relación romántica - consumo alcohol

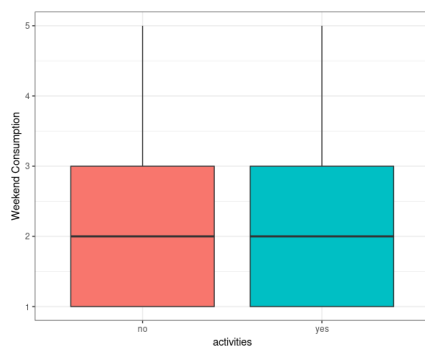


(a) Fin de semana

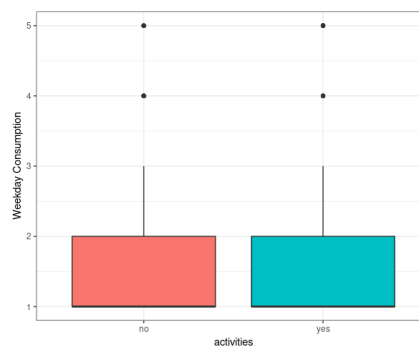


(b) Diario

Figura 17: Boxplot tutor - consumo alcohol

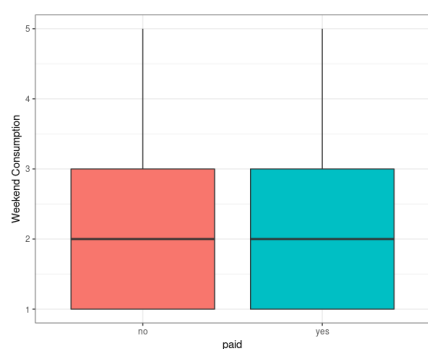


(a) Fin de semana

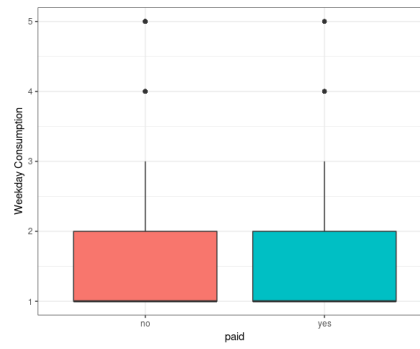


(b) Diario

Figura 18: Boxplot actividades extraescolares - consumo alcohol



(a) Fin de semana



(b) Diario

Figura 19: Boxplot clases pago - consumo alcohol

sex	age	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	traveltime
F:417	Min. :15.00	GT3:508	A: 90	0: 7	0: 7	at home :150	at home : 48	course :312	1:410
M:307	1st Qu.:16.00	LE3:216	T:634	1:165	1:199	health : 52	health : 26	home :171	2:237
	Median :17.00			2:205	2:235	other :283	other :407	other : 78	3: 59
	Mean :16.81			3:164	3:147	services:164	services:205	reputation:163	4: 18
	3rd Qu.:18.00			4:183	4:136	teacher : 75	teacher : 38		
	Max. :22.00								
failures	schoolsup	higher	famsup	famrel	freetime	goout	Dalc		
Min. :0.0000	no :648	no : 82	no :287	Min. :1.000	Min. :1.0	Min. :1.000	Min. :1.000		
1st Qu.:0.0000	yes: 76	yes:642	yes:437	1st Qu.:4.000	1st Qu.:3.0	1st Qu.:2.000	1st Qu.:1.000		
Median :0.0000				Median :4.000	Median :3.0	Median :3.000	Median :1.000		
Mean :0.3453				Mean :3.913	Mean :3.2	Mean :3.195	Mean :1.519		
3rd Qu.:0.0000				3rd Qu.:5.000	3rd Qu.:4.0	3rd Qu.:4.000	3rd Qu.:2.000		
Max. :3.0000				Max. :5.000	Max. :5.0	Max. :5.000	Max. :5.000		
Walc	health	studytime	absences	G					
Min. :1.000	Min. :1.000	1:236	Min. :0.00000	Min. : 3.823					
1st Qu.:1.000	1st Qu.:2.000	2:346	1st Qu.:0.03152	1st Qu.: 8.841					
Median :2.000	Median :4.000	3:106	Median :0.06791	Median :10.370					
Mean :2.311	Mean :3.552	4: 36	Mean :0.10154	Mean :10.657					
3rd Qu.:3.000	3rd Qu.:5.000		3rd Qu.:0.13387	3rd Qu.:12.333					
Max. :5.000	Max. :5.000		Max. :0.87333	Max. :18.333					

Figura 20: Summary dataset final

3.2. Estadística Inferencial

En este apartado vamos a ver si hay diferencia significativas en el consumo de alcohol a diario y fines de semana entre estudiantes en función del sexo y de situación de los padres. También veremos si el consumo de alcohol es distinto entre estudiantes que aprueban y que suspenden:

Se trata de un problema de diferencia de medias entre dos muestras en el que no conocemos la varianza poblacional. Tampoco sabemos a priori si los datos siguen una distribución normal, pero el tamaño de las muestras es lo suficientemente grande para tener en cuenta el teorema del límite central.

Para cada caso la hipótesis sería:

$$\begin{cases} H_0 : \mu_{g1} - \mu_{g2} = 0 \\ H_1 : \mu_{g1} - \mu_{g2} \neq 0 \end{cases}$$

3.2.1. Estudiantes de distinto sexo

Separamos el consumo de alcohol los fines de semana en dos conjunto según el sexo:

```
data_weekend_sex_m <- students$Walc[students$sex == 'M']
data_weekend_sex_f <- students$Walc[students$sex == 'F']

t.test(data_weekend_sex_m, data_weekend_sex_f, var.equal = TRUE,
conf.level = 0.95)
```

Two Sample t-test

```
data: data_weekend_sex_m and data_weekend_sex_f
t = 9.7324, df = 722, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 0.7114636 1.0710375
sample estimates:
mean of x mean of y
 2.824104  1.932854
```

Por el p-value vemos que no podemos aceptar la hipótesis nula, y concluimos que al 95 % de nivel de confianza los estudiantes masculinos y femeninos no

tienen el mismo consumo de alcohol los fines de semana.

Realizamos las mismas operaciones para el consumo de alcohol entre semana:

```
data_weekday_sex_m <- students$Dalc[students$sex == 'M']
data_weekday_sex_f <- students$Dalc[students$sex == 'F']

t.test(data_weekday_sex_m, data_weekday_sex_f, var.equal = TRUE,
conf.level = 0.95)
```

Two Sample t-test

```
data: data_weekday_sex_m and data_weekday_sex_f
t = 8.2948, df = 722, p-value = 5.321e-16
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 0.4211673 0.6823564
sample estimates:
mean of x mean of y
 1.837134  1.285372
```

Por el p-value vemos que ocurre lo mismo que en los fines de semana. No podemos aceptar la hipótesis nula y al 95 % afirmamos que hay diferencia en el consumo de alcohol entre estudiantes masculinos y femeninos para los días entre semana.

3.2.2. Estudiantes con diferente situación de convivencia de los padres

Ahora realizamos el estudio para el consumo de alcohol según la situación de los padres de cada alumno:

Primero, en fines de semana:

```
data_weekend_pstatus_t <- students$Walc[students$Pstatus == 'T']
data_weekend_pstatus_a <- students$Walc[students$Pstatus == 'A']

t.test(data_weekend_pstatus_t, data_weekend_pstatus_a, var.equal
= TRUE, conf.level = 0.95)
```

Two Sample t-test

```
data: data_weekend_pstatus_t and data_weekend_pstatus_a
t = 1.3035, df = 722, p-value = 0.1928
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 -0.09613914  0.47601996
sample estimates:
mean of x mean of y
 2.334385  2.144444
```

En el caso del consumo en los fines de semana, no hay diferencias significativas en función del estado de convivencia de los padres a un 95 % de nivel de confianza.

Ahora, para el consumo entre semana:

```
data_weekday_pstatus_t <- students$Dalc[students$Pstatus == 'T']
data_weekday_pstatus_a <- students$Dalc[students$Pstatus == 'A']

t.test(data_weekday_pstatus_t, data_weekday_pstatus_a, var.equal
= TRUE, conf.level = 0.95)
```

Two Sample t-test

```
data: data_weekday_pstatus_t and data_weekday_pstatus_a
t = 0.69872, df = 722, p-value = 0.4849
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 -0.1318160  0.2774872
sample estimates:
mean of x mean of y
 1.528391  1.455556
```

Con un p-value tan alto aceptamos la hipótesis nula, concluyendo que el consumo de alcohol entre semana es el mismo para estudiantes cuyos padres viven juntos y para estudiantes cuyos padres viven separados.

3.2.3. Estudiantes que aprueban ($G \geq 10$) y estudiantes que suspenden ($G < 10$)

Por último, realizamos el estudio para el consumo de alcohol según las calificaciones que obtiene el alumno (si aprueba o suspende):

En el caso del consumo los fines de semana:

```
data_weekend_aprobados <- students$Walc[students$G >= 10]
data_weekend_suspensos <- students$Walc[students$G < 10]

t.test(data_weekend_aprobados, data_weekend_suspensos, var.equal
= TRUE, conf.level = 0.95)
```

Two Sample t-test

```
data: data_weekend_aprobados and data_weekend_suspensos
t = -2.8183, df = 722, p-value = 0.00496
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 -0.4842420 -0.0865913
sample estimates:
mean of x mean of y
 2.214583  2.500000
```

Con un p-value menor que 0.05, al 95 % de confianza afirmamos que sí hay diferencia en el consumo de alcohol los fines de semana en función de si el alumno aprueba o no.

```
data_weekday_aprobados <- students$Dalc[students$G >= 10]
data_weekday_suspensos <- students$Dalc[students$G < 10]

t.test(data_weekday_aprobados, data_weekday_suspensos, var.equal
= TRUE, conf.level = 0.95)
```

Two Sample t-test

```
data: data_weekday_aprobados and data_weekday_suspensos
t = -2.7561, df = 722, p-value = 0.005996
```

```

alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 -0.34170360 -0.05740842
sample estimates:
mean of x mean of y
 1.452083  1.651639

```

Viendo el p-value, podemos decir al 95

3.3. Modelo de regresión lineal

En este apartado aplicaremos un modelo de regresión lineal múltiple que use como variables explicativas cuánto sale el estudiante con amigos, cuánto bebe entre semana, su sexo y su edad.

Al usar regresores cualitativos, es importante definir una categoría de referencia, para lo que usaremos la función de R `*relevel*` estableciendo la categoría "F" como referente para el sexo. El resultado lo almacenamos en una nueva variable.

```

students$sexR <- relevel(students$sex, "F")
modelo <- lm(Walc ~ goout + sexR + Dalc + age + studytime, data =
students )
summary(modelo)

```

Call:

```
lm(formula = Walc ~ G + sexR + Dalc + studytime + goout, data = students)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-3.3617	-0.6975	-0.1785	0.6505	2.8745

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.42554	0.18484	2.302	0.0216 *
G	-0.01084	0.01189	-0.912	0.3623
sexRM	0.39053	0.07592	5.144	3.48e-07 ***
Dalc	0.69001	0.04087	16.885	< 2e-16 ***
studytime2	-0.14179	0.08172	-1.735	0.0832 .

```

studytime3  -0.25165    0.11580   -2.173    0.0301 *
studytime4  -0.43120    0.16952   -2.544    0.0112 *
goout        0.28685    0.03076    9.327   < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9339 on 716 degrees of freedom
Multiple R-squared:  0.4843, Adjusted R-squared:  0.4793
F-statistic: 96.06 on 7 and 716 DF,  p-value: < 2.2e-16

```

Como podemos ver, aspectos que influyen mucho en el modelo son:

- Que el alumno salga con amigos
- Que el alumno sea de sexo masculino
- Si el alumno bebe entre semana tenderá a beber más los fines de semana
- Cuanto más tiempo de estudio dedica el alumno, menos bebe

La edad no influye de forma significativa para el consumo de alcohol según el modelo.

En este caso, el modelo explica un 48 % de la variabilidad en los datos.

4. Estudio de correlación

Estudiamos el nivel de significancia de la relación entre la calificación que obtienen los alumnos y otro tipo de factores:

- Sexo
- Edad
- Consumo de alcohol los fines de semana
- Consumo de alcohol entre semana
- Situación de los padres (cohabitando o no)

Al estudiar el nivel de relación entre una variable continua (la nota) y variables categóricas (el resto), usamos el test ANOVA para obtener este nivel de significancia:

```
# Sexo
aov1 = aov(students.red$G ~ students.red$sex)
summary(aov1)
# Edad
aov1 = aov(students.red$G ~ students.red$age)
summary(aov1)
# Consumo de alcohol fines de semana
aov1 = aov(students.red$G ~ students.red$Walc)
summary(aov1)
# Consumo de alcohol entre semana
aov1 = aov(students.red$G ~ students.red$Dalc)
summary(aov1)
# Situación de los padres
aov1 = aov(students.red$G ~ students.red$Pstatus)
summary(aov1)
```

Output:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
students.red\sex	1	24	24.46	2.688	0.102
Residuals	722	6570	9.10		

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
students.red\$age	1	109	109.37	12.18	0.000513 ***
Residuals	722	6485	8.98		

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

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
students.red\$Walc	1	150	150.35	16.84	4.52e-05 ***
Residuals	722	6444	8.93		

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

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
students.red\$Dalc	1	139	139.34	15.58	8.66e-05 ***
Residuals	722	6455	8.94		

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

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
students.red\$Pstatus	1	1	1.231	0.135	0.714
Residuals	722	6594	9.132		

Como podemos observar, no hay correlación entre la calificación del estudiante y el sexo o el estado de convivencia de los padres. Sin embargo, en el desempeño escolar de los estudiantes sí que influyen significativamente la edad y el consumo de alcohol tanto en fines de semana como entre semana.

5. Conclusiones

POR COMPLETAR

Referencias

- [HKP12] HAN, Jiawei ; KAMBER, Micheline ; PEI, Jian: *Data mining: concepts and techniques*. Elsevier/Morgan Kaufmann, 2012. – Chapter 3
- [Osb10] OSBORNE, Jason W.: Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. In: *Newborn and Infant Nursing Reviews* 10 (2010), Nr. 1, S. 37–43. <http://dx.doi.org/10.1053/j.nainr.2009.12.009>. – DOI 10.1053/j.nainr.2009.12.009
- [Squ15] SQUIRE, Megan: *Clean data: save time by discovering effortless strategies for cleaning, organizing, and manipulating your data*. Packt Publishing Ltd, 2015. – Chapters 1 & 2