Estudio retrospectivo sobre las capacidades en matemáticas de los estudiantes españoles en el examen PISA 2012

PISA

Programme for International Student Assessment

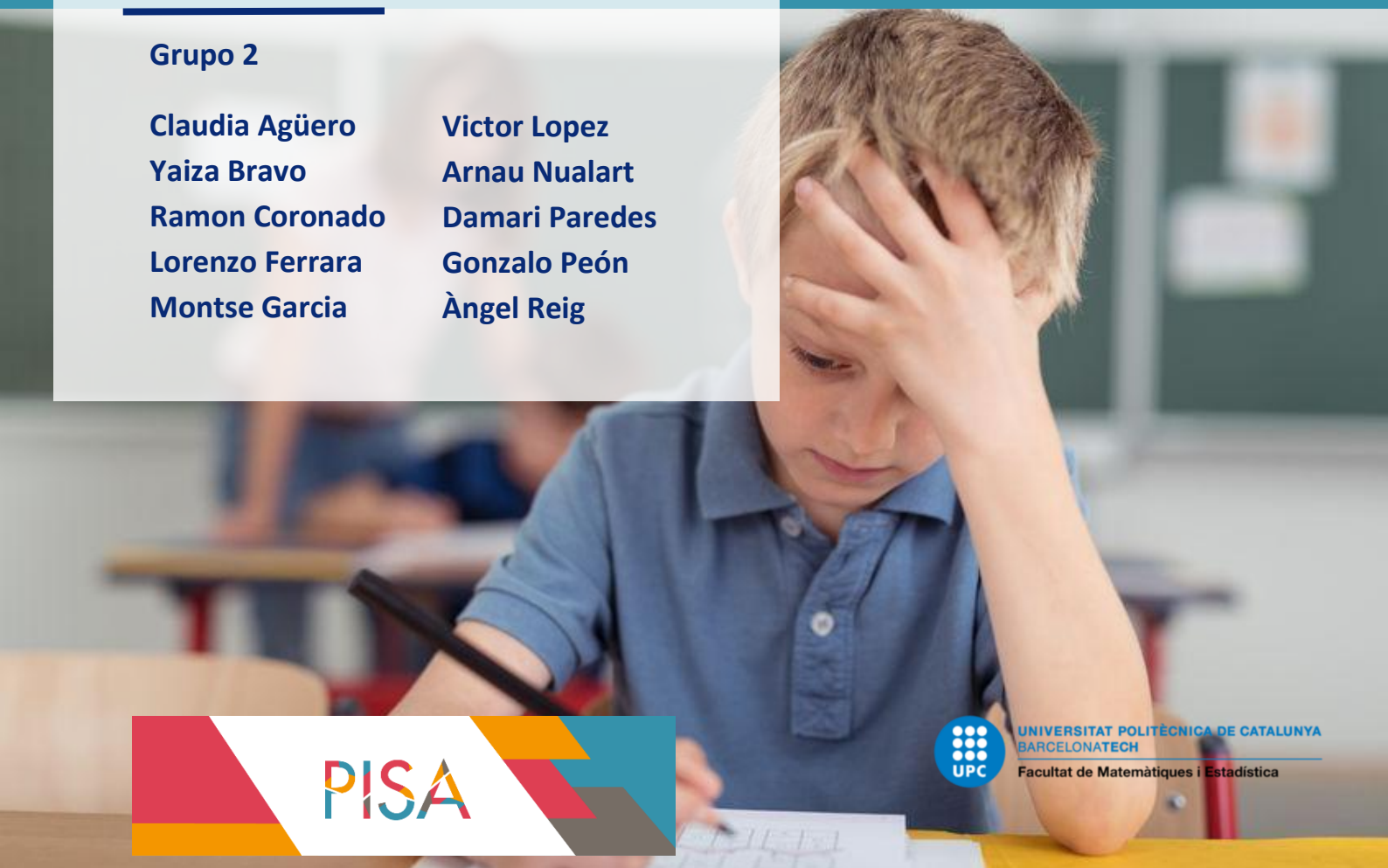
OECD

BETTER POLICIES FOR BETTER LIVES

Entrega D3

Grupo 2

Claudia Agüero
Yaiza Bravo
Ramon Coronado
Lorenzo Ferrara
Montse Garcia
Victor Lopez
Arnau Nualart
Damari Paredes
Gonzalo Peón
Àngel Reig



PISA

UPC
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Tabla de contenidos

Bibliografia

- **Base de Datos**
  https://www.oecd.org/pisa/pisaproducts/pisa2012database-downloadabledata.htm
- **Codebook**
  https://www.oecd.org/pisa/pisaproducts/PISA12_stu_codebook.pdf
- **Repositorio Drive**
  https://drive.google.com/drive/u/0/folders/1yyOO2In_Ek5SPqdSVQxoEHuPNMFBC5eL

Descripción del dataset

PISA es el Programa para la Evaluación Internacional de Estudiantes de la OCDE. Es una prueba estandarizada a nivel internacional compuesta por un cuestionario inicial y una serie de ejercicios para observar y medir los conocimientos de jovenes de 15 años en materias como la lectura, matemáticas y ciencias. La base de datos del proyecto fue obtenida de la prueba realizada en 2012 para España. Se han estudiado los resultados de los alumnos españoles en la parte de matemáticas en base a una serie de variables explicativas con el fin de predecir la nota de su examen.

La base de datos utilizada es “pisa.csv”, esta base se ha extraído de una base más grande llamada “data_esp.sav”, la cual contiene 25.313 elementos de 14 variables distintas. De estos datos se extrae de forma aleatoria nuestra muestra, que representa solamente un 50% de todas las observaciones.

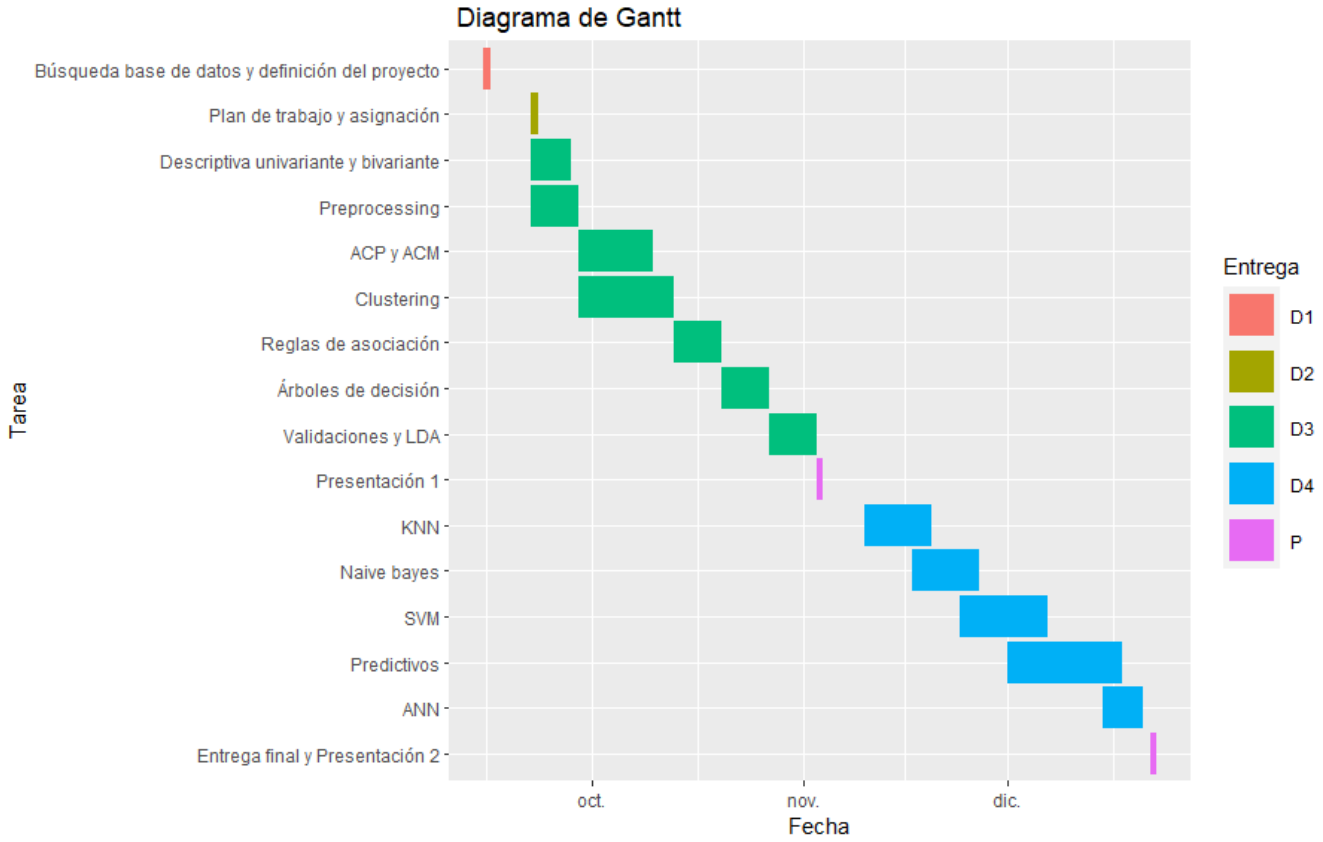
Codebook

Variable	Descripción	Tipo	Rango/Categorías	Rol	Missing
School	Identificador de la escuela	Categórica	0000001 - 0000902	Explicatoria	*"NA"*
Type	Tipo de Escuela	Binaria	Public / Private	Explicatoria	*"NA"*
CCAA	Subregión (Comunidad Autónoma) de la escuela	Categórica	Andalusia, Basque Country, Catalonia, Extremadura, Galicia, La Rioja, Madrid, Murcia …Other	Explicatoria	*"Unknown"*
Student	ID del estudiante (único dentro de la escuela)	Categórica	00003, 00004, 25309	Explicatoria	*"Missing"*
Sex	Sexo del estudiante	Binaria	Male / Female	Explicatoria	*"NA"*
Age	Edad del estudiante	Númerica	[15.33, 16.33]	Explicatoria	*"*"*
Repeated	Si el estudiante repitió curso en secundaria	Categórica	Never, Once, Twice or more, N/A, Invalid, Missing	Respuesta	*"Missing"*
Homework	Frecuencia en la que el estudiante realiza deberes fuera de clase	Categórica	Never or hardly ever, Once/Twice a month, Once/Twice a week, Almost every day, Every day	Explicatoria	*"Unknown"*

Teach_int	Frecuencia con la cual el docente de matemáticas muestra interés	Categórica	Never or hardly ever, Some Lessons, Most Lessons, Every Lesson	Explicatoria	*"Unknown"*
Score	Puntuación esperada en matemáticas una vez corregidos los efectos de booklet y de las dificultades de las preguntas	Númerica	[100, 800]	Explicatoria	*"NA"*
Belief	Media Likert del conocimiento que el estudiante cree tener sobre 10 campos de las matemáticas	Númerica	[-5.16, 5.18]	Explicatoria	*"*"*
Foil	Media Likert del conocimiento que el estudiante cree tener sobre 3 campos de las matemáticas no existentes	Númerica	[-5, 5]	Explicatoria	*"*"*
St_math	Minutos que el estudiante dedica a la semana a estudiar matemáticas	Númerica	[0,500]	Explicatoria	*"NA"*
St_par	Horas semanales de estudio con al menos uno de los padres	Númerica	[0,30]	Explicatoria	*"NA"*
Par_ed	Años máximos de educación alcanzado por los padres	Númerica	[3,16.5]	Explicatoria	*"NA"*

Diagrama de Gantt

A continuación, se presenta el timeline del proyecto a través del diagrama de Gantt. Las actividades se agrupan por colores, según el calendario de entregas. Como vemos, la mayor parte del proyecto se concentra en las actividades de procesamiento de datos y en la predicción de resultado. La última actividad finaliza el día de entrega del trabajo, con su presentación, el 22 de Diciembre.



Plan de Riesgos

Posibles inconvenientes durante la realización del proyecto. Se proponen medidas de prevención y soluciones en caso de ocurrencia.

RIESGO	CÓMO EVITARLO	SOLUCIÓN
Ausencia de un miembro del equipo	Seguir el programa de tareas	Reasignar sus tareas al resto del grupo
Perder la coordinación entre los miembros del grupo	Respetar el diagrama de Gantt y la repartición de tareas	Evitar los archivos duplicados
Renuncia de la evaluación continua por parte de algún miembro	Es un riesgo inevitable	Reorganizar tareas
Pasar un plazo de entrega	Seguir el programa de tareas	Adaptar el diagrama de Gantt
No saber realizar una tarea	Ir al día con el temario de la asignatura	Preguntar al tutor o a otro miembro del grupo
Falta de material o herramientas informáticas	Prever que será necesario para cada tarea	Recurrir a los medios que proporcione la universidad
Un compañero deja el curso	Todas las tareas tienen 3 o 4 miembros asignados	Los otros compañeros asumen su parte

Peligrosidad: Grave Moderada Baja

Distribución de Tareas

	Claudia	Yaiza	Ramon	Lorenzo	Montse	Victor	Arnau	Damari	Gonzalo	Àngel
Búsqueda base de datos y definición del proyecto			x	x				x	x	
Plan de trabajo y asignación		x			x	x				
Descriptiva univariante y bivariante	x						x	x		x
Preprocessing			x	x					x	
ACP y ACM	x	x				x				
Clustering					x		x			x
Reglas de asociación			x			x			x	
Árboles de decisión	x			x	x			x		
Validaciones y LDA		x	x				x			x
KNN	x			x		x			x	
Naive bayes			x		x			x		x
SVM		x		x			x			
Predictivos	x					x		x		x
ANN		x			x		x		x	

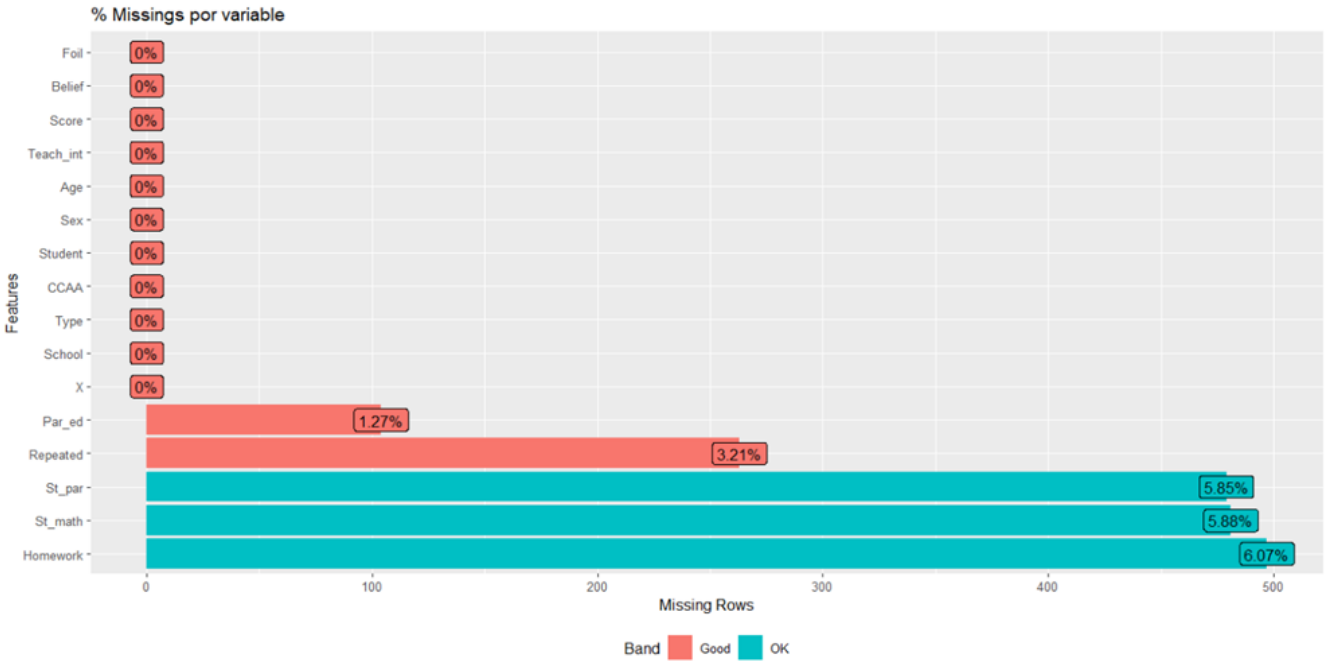
Descriptiva Univariante Antes del Preprocessing

Para empezar, debemos tener un primer contacto con nuestra base de datos.

Missings

En la Figura 1 podemos ver el porcentaje de missings que presenta cada variable. Podemos observar que la mayoría de nuestras variables contienen datos completos. Las variables con valores más bajos de missings son Par_ed (1.27%) y Repeated (3.21%). Las variables que contienen más missings son tan solo 3 (St_par, St_math y Homework) y muestran cerca de un 6% de missings.

Figura 1: Missings de las variables



Variables Numéricas

Estadísticos principales

A continuación, en la Tabla 1 y Tabla 2 podemos observar la tabla de los estadísticos principales para cada variable numérica.
La variable X es un indicador que no tendremos en cuenta a la hora de hacer el análisis a pesar de verla aquí. También, hay que considerar que las variables School y Student deberán ser tratadas posteriormente como categóricas.

Tabla 1: Estadísticos variables numéricas

Variables	n_missing	complete_rate	mean	sd
X	0	1.00	4094.00	2363.53
School	0	1.00	451.24	261.38
Student	0	1.00	12692.14	7314.30
Age	0	1.00	15.87	0.29
Score	0	1.00	498.14	84.65
Belief	0	1.00	0.81	1.21
Foil	0	1.00	-0.52	1.17
St_math	481	0.94	206.96	37.60
St_par	479	0.94	0.88	2.22
Par_ed	104	0.99	12.84	3.55

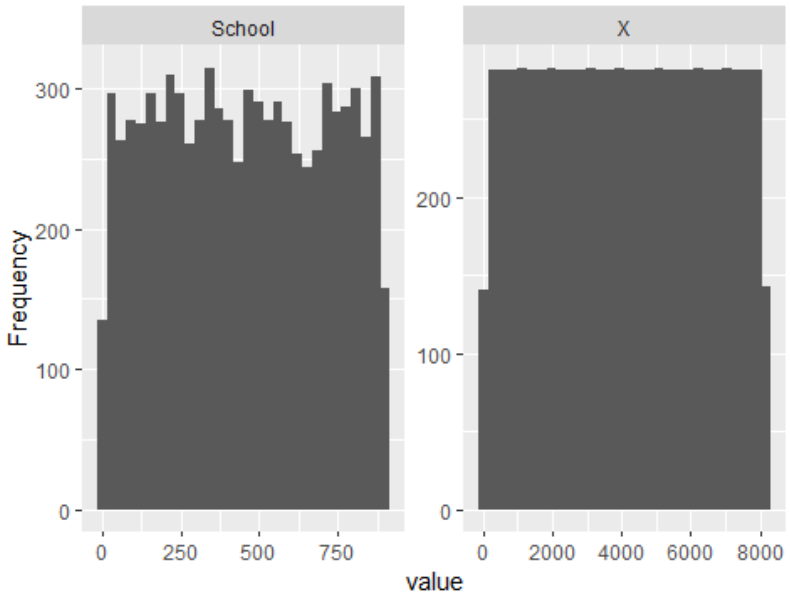
Tabla 2: Estadísticos variables numéricas

Variables	mín	p25	p50	p75	máx
X	1.00	2047.50	4094.00	6140.50	8187.00
School	1.00	225.00	451.00	684.00	902.00
Student	3.00	6398.50	12701.00	19048.00	25309.00
Age	15.33	15.58	15.92	16.08	16.33
Score	130.06	441.04	500.40	558.86	763.13
Belief	-5.16	-0.10	0.96	1.68	5.18
Foil	-2.31	-1.47	-0.80	0.30	4.24
St_math	135.00	180.00	200.00	220.00	500.00
St_par	0.00	0.00	0.00	1.00	30.00

Par_ed	3.00	12.00	13.00	16.50	16.50

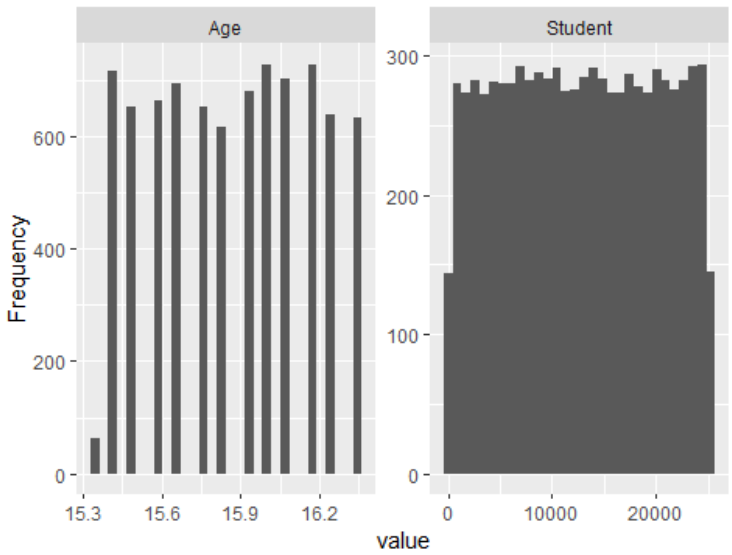
Histogramas

Histogramas 1: School y X



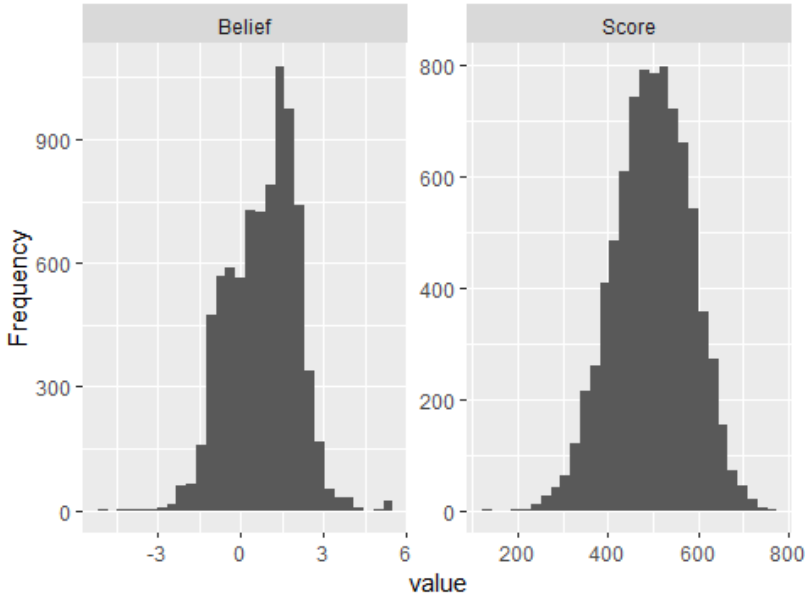
X es una variable indicadora que no utilizaremos. School y Student son variables con patrones muy similares, ya que sus valores son indicadores que posteriormente deberán ser considerados como factores. No observamos nada muy destacable, ya que a pesar de haber escuelas con más o menos estudiantes, no se aprecia ningún valor muy extremo.

Histogramas 2: Age y Student



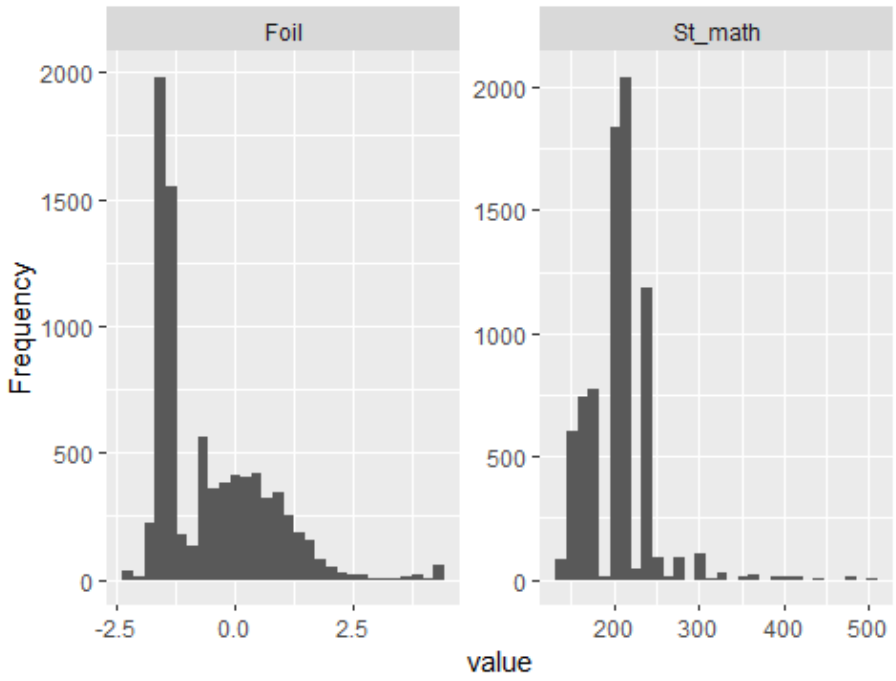
Por lo que respecta a la variable Age, vemos que se encuentra en el intervalo establecido dentro de los parámetros que dictaba la prueba PISA. Lo único que podemos destacar es que el grupo de alumnos de edades cercanas a los 15 años exactos es más pequeño que el resto.

Histogramas 3: Belief y Score



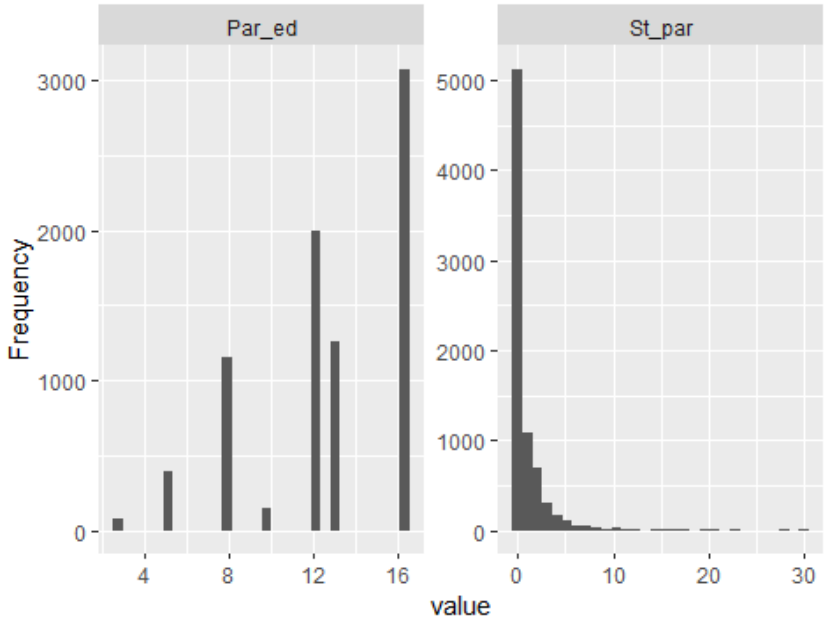
El histograma de Score parece acorde a una distribución normal. Y Belief podría también serlo pero está un poco menos claro. Tiene una ligera inclinación hacia el lado derecho, lo cúal podría indicar que los alumnos creen tener un conocimiento medio sobre 10 campos de las matemáticas.

Histogramas 4: Foil y St_math



Tanto para Foil como para St_math observamos que los datos se acumulan hacia los valores más bajos. Y muy pocos datos en los extremos positivos.

Histogramas 5: Par_ed y St_par



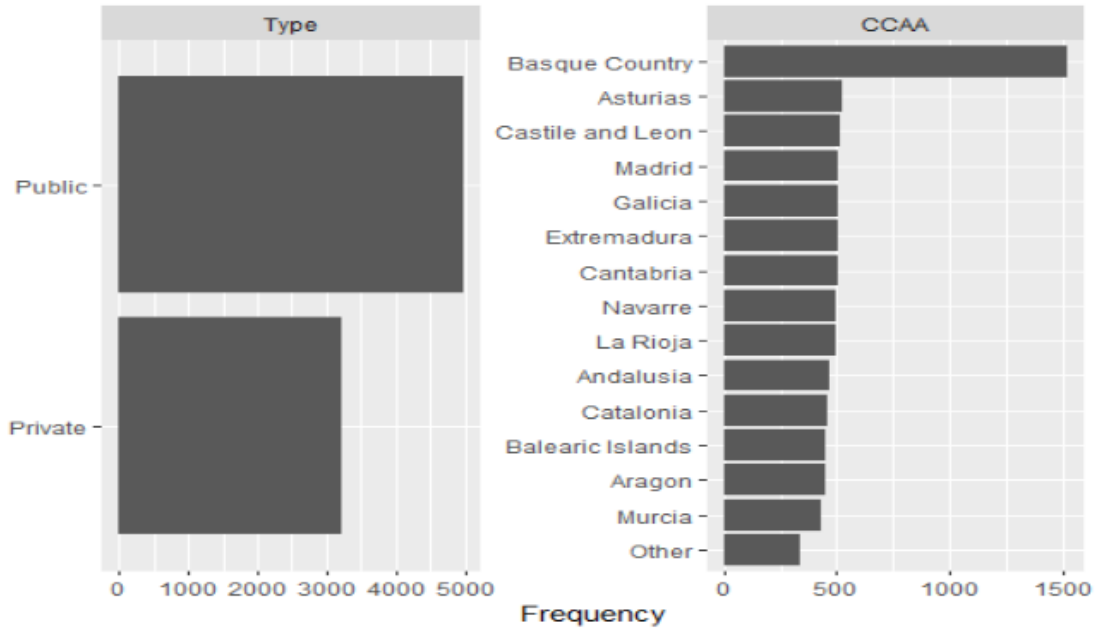
En el caso de Par_ed observamos que hay una tendencia creciente a medida que los valores son más elevados (hay más alumnos con padres con unos 16 años de educación alcanzados). Por otro lado, observamos que la mayoría de alumnos no estudian con sus padres (St_par).

Variables Categóricas

A continuación, hacemos una descripción de las variables categóricas, mediante tablas de frecuencias.

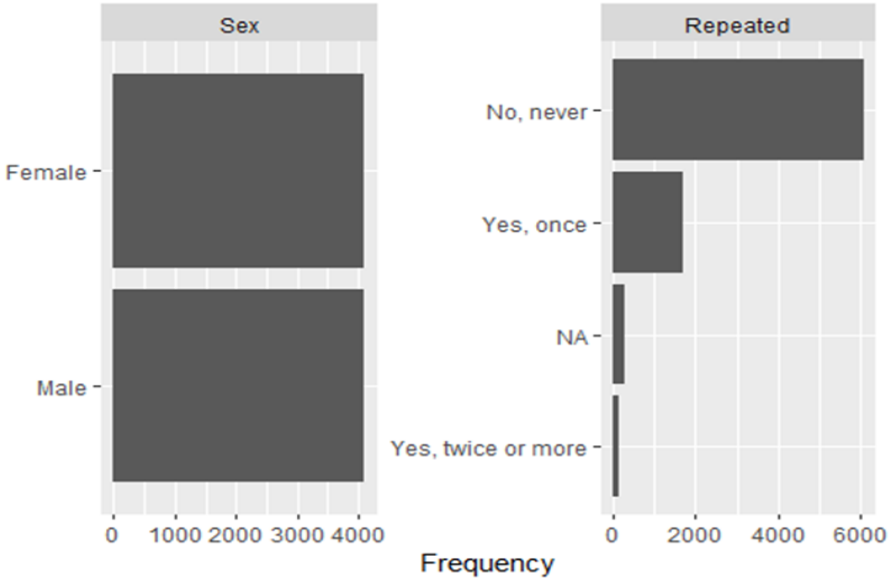
Respecto a Type (tipo de escuela), podemos ver que predominan las escuelas privadas. Y por lo que podemos observar de las comunidades autónomas, la mayoría de alumnos provienen del País Vasco, muy por encima del resto que tienen una frecuencia más similar entre ellas.

Tabla de Frecuencias 1: Type y CCAA



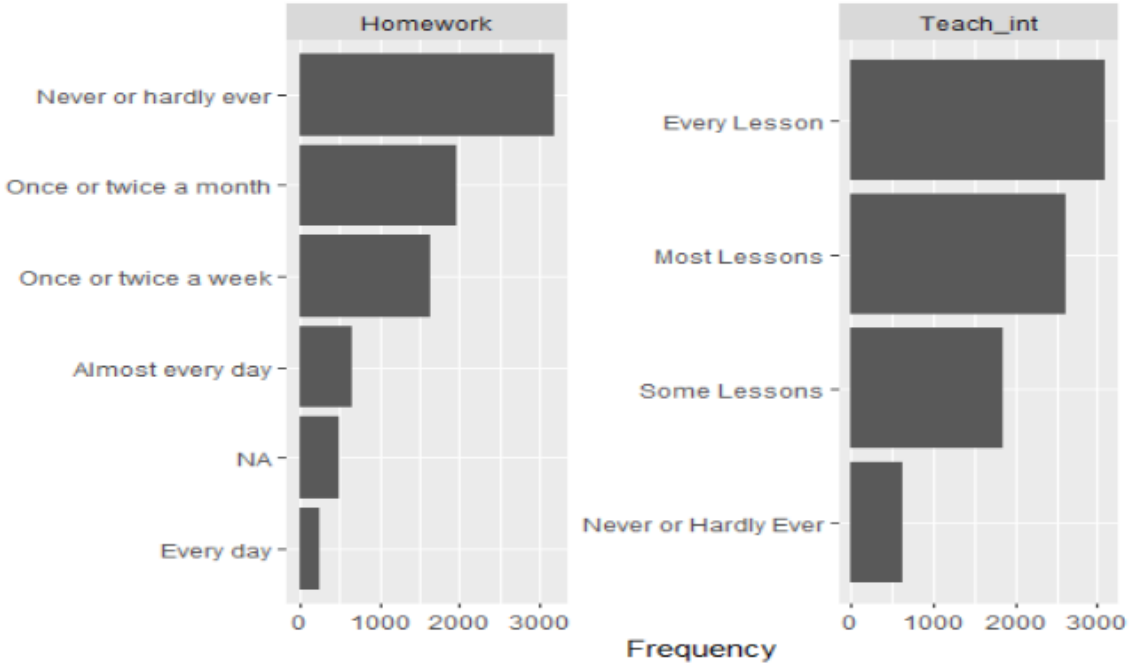
En cuanto a género tenemos la misma cantidad de chicos que de chicas. Mientras que predominan los alumnos que no han repetido curso.

Tabla de Frecuencias 2: Sex y Repeated



Vemos en la tabla de frecuencias 3, que el alumnado no suele hacer los deberes fuera de la escuela y que a medida que aumenta esta frecuencia, disminuye el número de alumnos. Por otro lado, el número de docentes de matemáticas que muestran interés en cada lección es mayoritario. Siendo pocos los que nunca muestran interés.

Tabla de Frecuencias 3: Homework y Teach_int

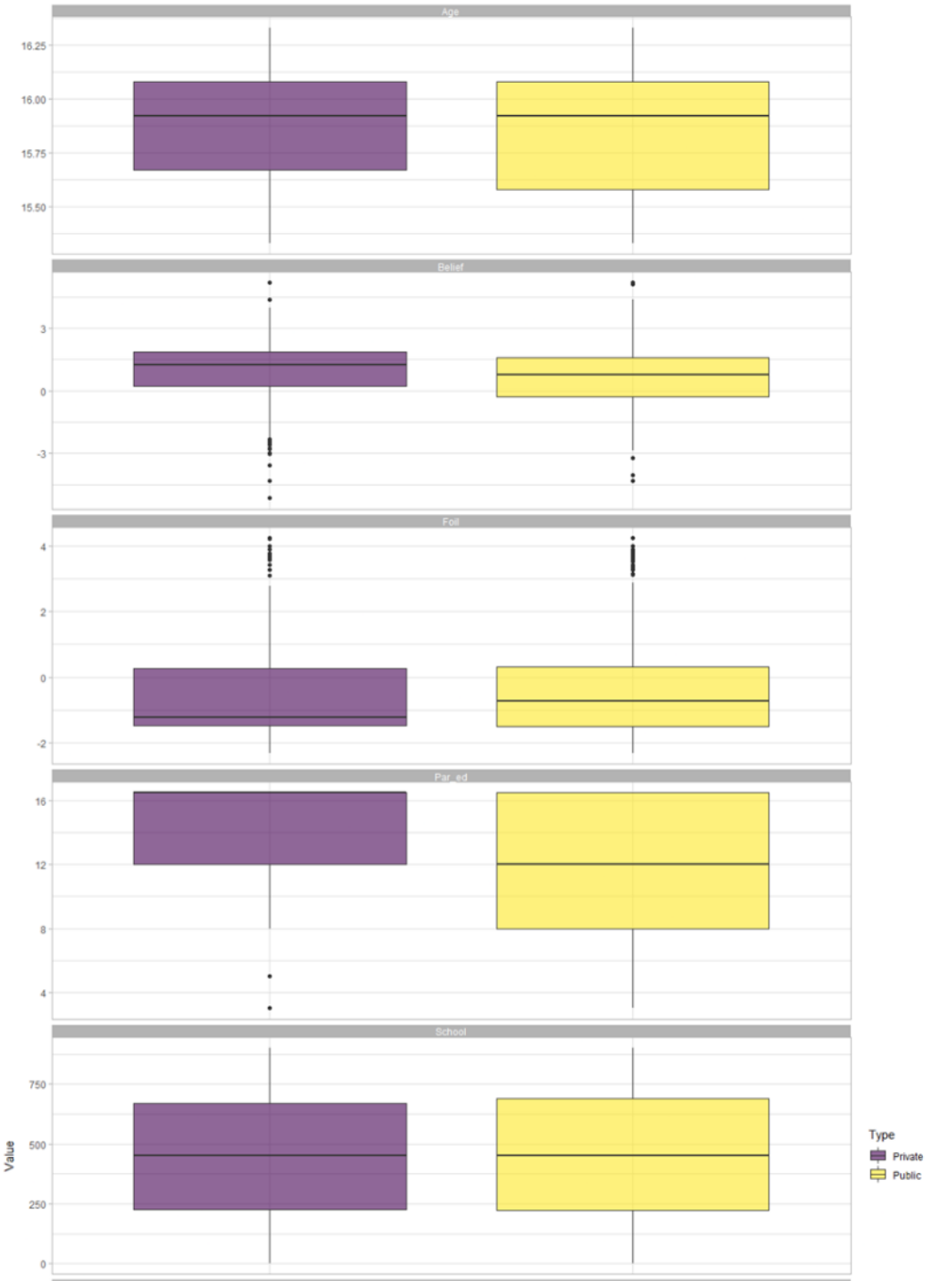


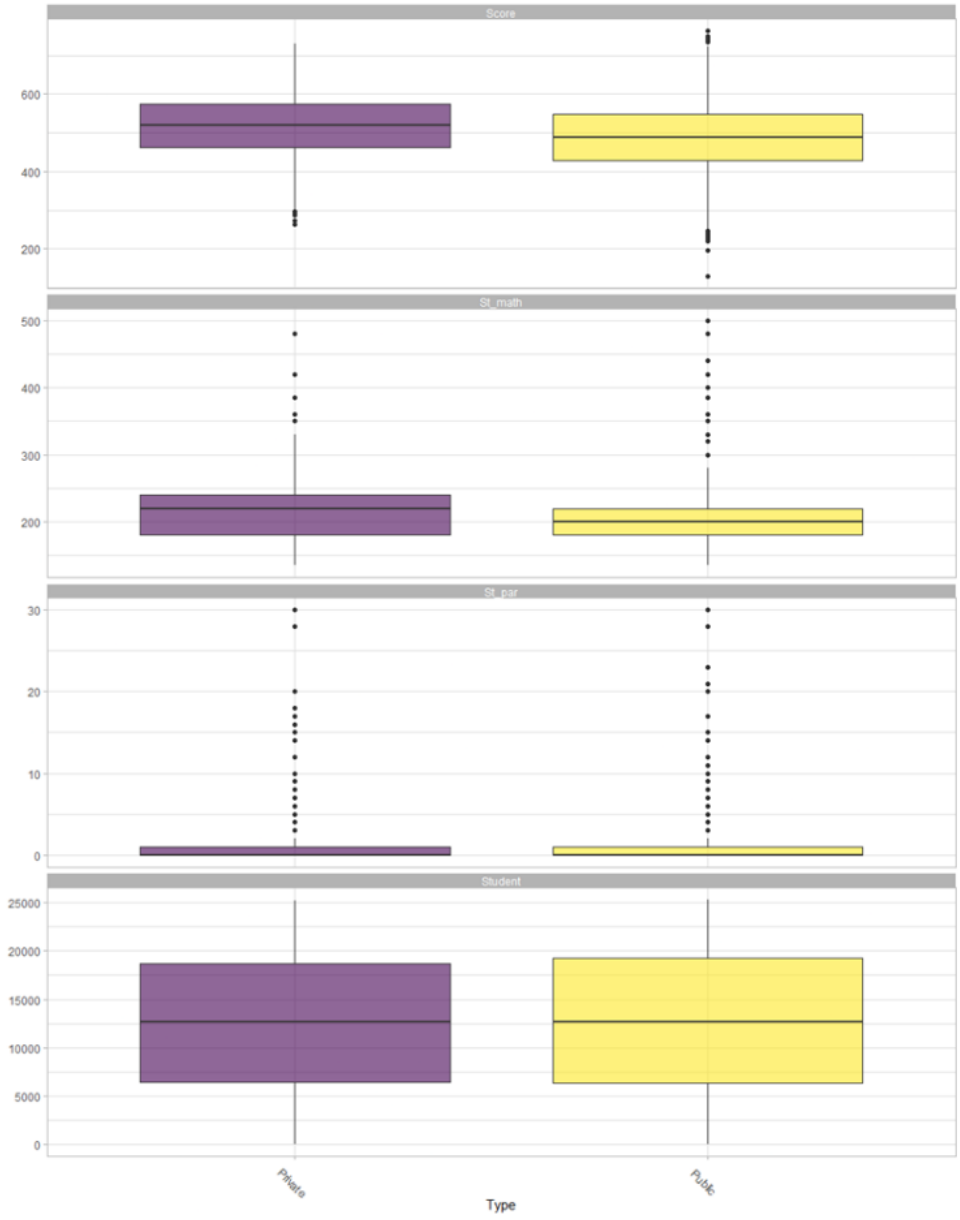
Descriptiva Bivariante Antes del Preprocessing

Para empezar debemos tener un primer contacto con nuestra base de datos. A continuación analizaremos la variable type (nuestra variable respuesta) con las distintas variables numericas.

Descriptiva Bivariante Categórica-Numérica (Pre)

Type


Age
16.25
16.00
15.75
15.50
Belief
3
0
-3
Fail
4
2
0
-2
Par_ed
16
12
8
4
School
750
500
250
0
Value
Type
Private
Public


Score
600
400
200
St_math
500
400
300
200
St_par
30
20
10
0
Student
25000
20000
15000
10000
5000
0
Private
Public
Type

En la tabla a continuación podemos ver los estadísticos principales de la nuestra variable respuesta Type(Categórica) con cada una de las variables numéricas.

Tabla 4: Estadísticos variables numéricas

variable	Type	N	Mean	SD	Min	Max	Median	Q1	Q3	p.value
School	Private	3,223	448.142	263.173	2.000	900.000	450.000	225.000	669.500	0.3876
School	Public	4,964	453.251	260.222	1.000	902.000	451.000	223.000	691.000	
Student	Private	3,223	12,608.186	7,368.436	37.000	25,250.000	12,684.000	6,409.500	18,728.500	0.4027
Student	Public	4,964	12,746.655	7,279.161	3.000	25,309.000	12,704.500	6,339.750	19,243.750	
Age	Private	3,223	15.874	0.288	15.330	16.330	15.920	15.670	16.080	0.2414
Age	Public	4,964	15.867	0.290	15.330	16.330	15.920	15.580	16.080	
Score	Private	3,223	516.690	78.957	263.522	730.885	520.027	462.852	573.773	< 0.0001
Score	Public	4,964	486.100	86.050	130.059	763.133	487.545	427.313	547.620	
Belief	Private	3,223	1.036	1.195	-5.160	5.180	1.250	0.200	1.850	< 0.0001
Belief	Public	4,964	0.666	1.198	-4.320	5.180	0.770	-0.275	1.580	
Foil	Private	3,223	-0.561	1.166	-2.312	4.244	-1.218	-1.470	0.278	0.0189
Foil	Public	4,964	-0.499	1.165	-2.312	4.244	-0.715	-1.487	0.309	
St_math	Private	3,080	211.245	34.033	135.000	480.000	220.000	180.000	240.000	< 0.0001
St_math	Public	4,626	204.110	39.549	135.000	500.000	200.000	180.000	220.000	
St_par	Private	3,062	0.825	2.088	0.000	30.000	0.000	0.000	1.000	0.0681
St_par	Public	4,646	0.920	2.300	0.000	30.000	0.000	0.000	1.000	
Par_ed	Private	3,196	13.848	3.123	3.000	16.500	16.500	12.000	16.500	< 0.0001
Par_ed	Public	4,887	12.188	3.657	3.000	16.500	12.000	8.000	16.500	

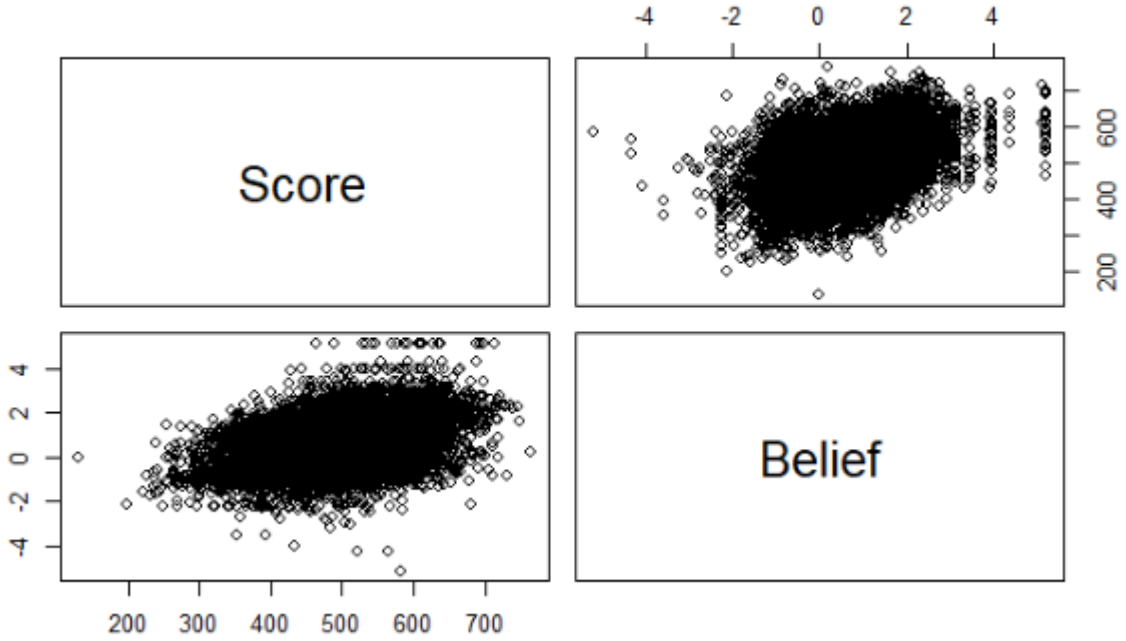
Según el test de Kruskal-Wallis, existen diferencias significativas entre las diferentes veces que el estudiante ha repetido curso (Repeated) de: Edat, Score, Belief, St_math, St_par y Par_ed.

Descriptiva Bivariante Numérica-Numérica

Calculamos las correlaciones entre variables y miramos cuál es la correlación máxima para encontrar las variables con mayor correlación. Debido a que todavía nos encontramos antes del preprocessing vemos que existen varias variables con NA's.

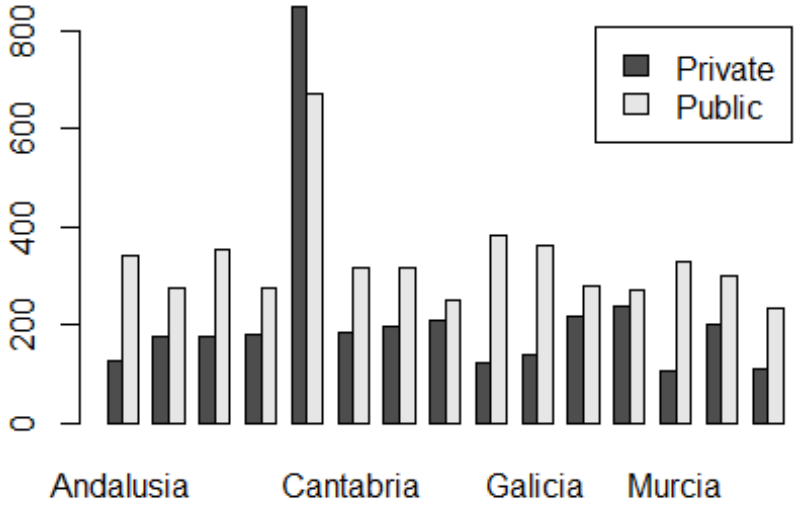
	School	Student	Age	Score	Belief	Foil	St_math	St_par	Par_ed
School	1.000000000	0.999951314	0.002379014	-0.007354498	0.021782164	-0.01918652	NA	NA	NA
Student	0.999951314	1.000000000	0.002363755	-0.007461227	0.021771870	-0.01942877	NA	NA	NA
Age	0.002379014	0.002363755	1.000000000	0.043825447	0.003685431	0.03975067	NA	NA	NA
Score	-0.007354498	-0.007461227	0.043825447	1.000000000	0.408644957	-0.01609374	NA	NA	NA
Belief	0.021782164	0.021771870	0.003685431	0.408644957	1.000000000	-0.64361687	NA	NA	NA
Foil	-0.019186520	-0.019428768	0.039750666	-0.016093745	-0.643616871	1.00000000	NA	NA	NA
St_math	NA	NA	NA	NA	NA	NA	1	NA	NA
St_par	NA	NA	NA	NA	NA	NA	NA	1	NA
Par_ed	NA	NA	NA	NA	NA	NA	NA	NA	1

La correlación más extrema corresponde a Score y Belief ya que la correlación entre School y Student no es válida debido a que ambas son variables identificativas. Esta tiene un valor de -0.40, la mayoría no estan muy correlacionadas.



Descriptiva Bivariante Categórica-Categórica

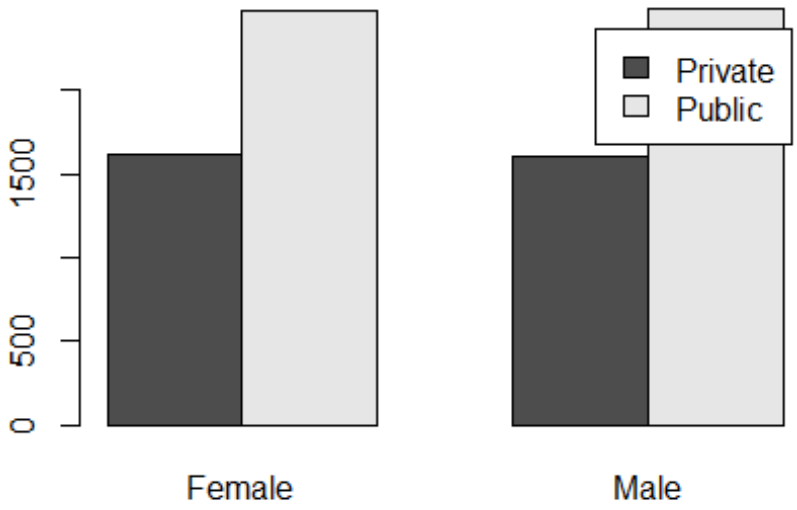
CCAA




```
##    Cell Contents
## |-------------------------|
## |                   Count |
## |         Expected Values |
## |-------------------------|
##
## =========================================
##                     ddcat$Type
## ddcat[[k]]          Private   Public   Total
## -----------------------------------------
## Andalusia               128      343     471
##                         185      286
## -----------------------------------------
## Aragon                  177      274     451
##                         178      274
## -----------------------------------------
## Asturias                174      355     529
##                         208      321
```

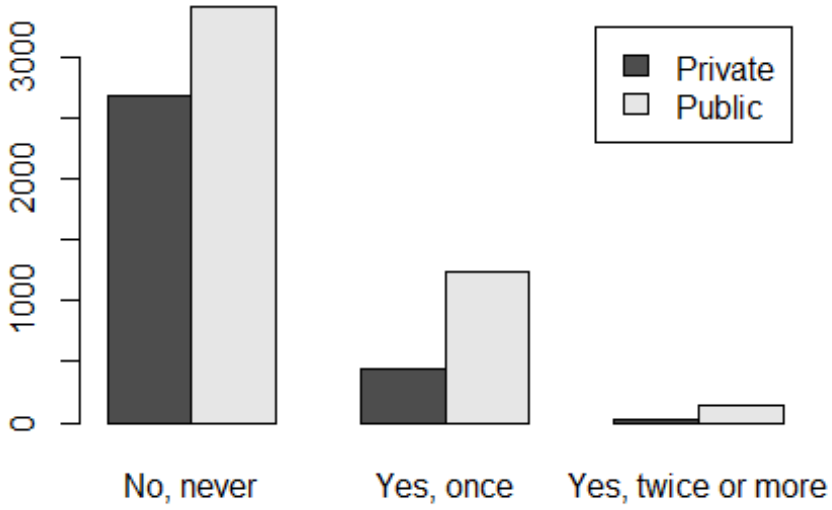
```
## ------------------------------------------
## Balearic Islands        178      275     453
##                         178      275
## ------------------------------------------
## Basque Country          849      674    1523
##                         600      923
## ------------------------------------------
## Cantabria               186      316     502
##                         198      304
## ------------------------------------------
## Castile and Leon        197      316     513
##                         202      311
## ------------------------------------------
## Catalonia               210      252     462
##                         182      280
## ------------------------------------------
## Extremadura             121      381     502
##                         198      304
## ------------------------------------------
## Galicia                 138      364     502
##                         198      304
## ------------------------------------------
## La Rioja                217      280     497
##                         196      301
## ------------------------------------------
## Madrid                  237      271     508
##                         200      308
## ------------------------------------------
## Murcia                  104      329     433
##                         170      262
## ------------------------------------------
## Navarre                 199      301     500
##                         197      303
## ------------------------------------------
## Other                   108      233     341
##                         134      207
## ------------------------------------------
## Total                  3223     4964    8187
## ==========================================
##
## Statistics for All Table Factors
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 = 363.314      d.f. = 14      p <2e-16
##
##         Minimum expected frequency: 134.2425
```

SEX



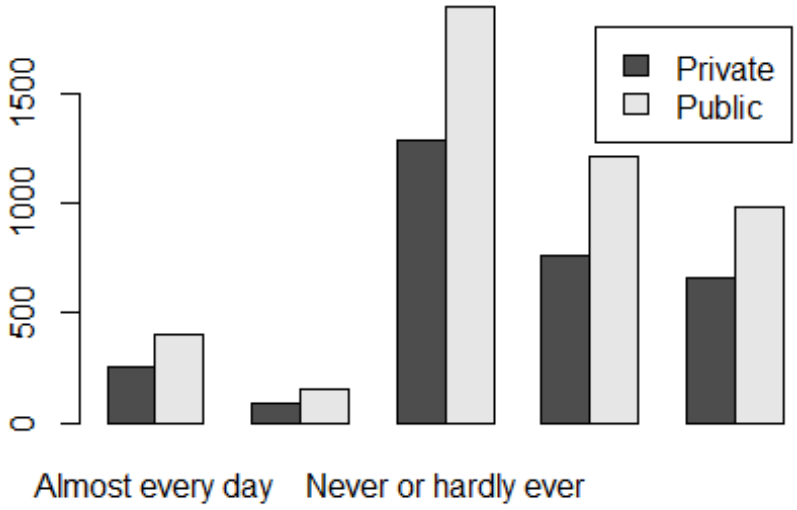

```
##    Cell Contents
## |-------------------------|
## |                   Count |
## |         Expected Values |
## |-------------------------|
##
## ====================================
##               ddcat[[k]]
## ddcat$Type    Female   Male   Total
## ------------------------------------
## Private         1622   1601    3223
##                 1614   1608
## ------------------------------------
## Public          2479   2485    4964
##                 2486   2478
## ------------------------------------
## Total           4101   4086    8187
## ====================================
##
## Statistics for All Table Factors
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 = 0.1165991      d.f. = 1      p = 0.733
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 = 0.101662      d.f. = 1      p = 0.75
##         Minimum expected frequency: 1608.547
```

REPEATED




```
##    Cell Contents
## |-------------------------|
## |                   Count |
## |         Expected Values |
## |-------------------------|
##
## ================================================================
##               ddcat[[k]]
## ddcat$Type    No, never   Yes, once   Yes, twice or more   Total
## ----------------------------------------------------------------
## Private            2675         442                   28    3145
##                    2413         668                   64
## ----------------------------------------------------------------
## Public             3405        1240                  134    4779
##                    3667        1014                   98
## ----------------------------------------------------------------
## Total              6080        1682                  162    7924
## ================================================================
##
## Statistics for All Table Factors
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 = 207.4825      d.f. = 2      p <2e-16
##
##         Minimum expected frequency: 64.29707
```

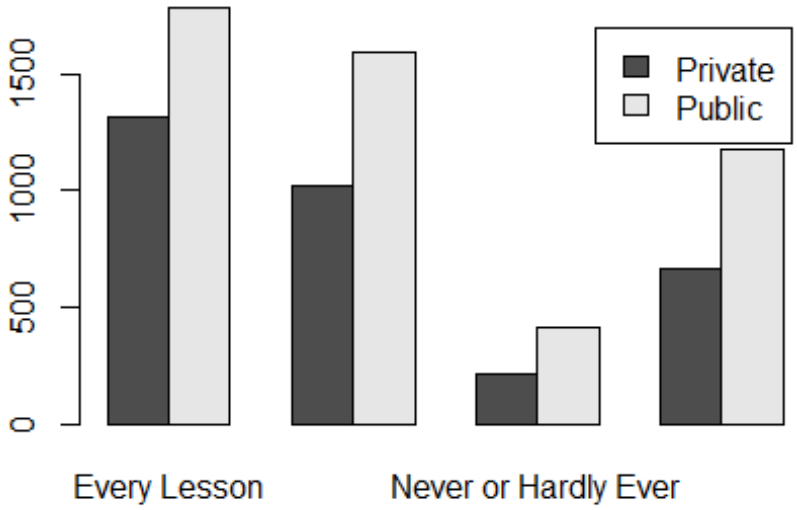
HOMEWORK




```
##    Cell Contents
## |-------------------------|
## |                   Count |
## |         Expected Values |
## |-------------------------|
##
##
=============================================================================
##              ddcat[[k]]
## ddcat$Typ    Almst e d   Every day   Nvr o h e   O o t a m   O o t a w
Total
## -----------------------------------------------------------------------------
--
## Private            254          86        1290         761         658
3049
##                    259          96        1262         781         651
## -----------------------------------------------------------------------------
--
## Public             400         156        1893        1209         983
4641
##                    395         146        1921        1189         990
## -----------------------------------------------------------------------------
--
## Total              654         242        3183        1970        1641
7690
##
=============================================================================
##
## Statistics for All Table Factors
```

```
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 = 3.910747      d.f. = 4      p = 0.418
##
##         Minimum expected frequency: 95.95033
```

TEACH_INT




```
##    Cell Contents
## |-------------------------|
## |                   Count |
## |         Expected Values |
## |-------------------------|
##
## ================================================
##                          ddcat$Type
## ddcat[[k]]               Private   Public   Total
## ------------------------------------------------
## Every Lesson                1320     1784    3104
##                             1222     1882
## ------------------------------------------------
## Most Lessons                1018     1595    2613
##                             1029     1584
## ------------------------------------------------
## Never or Hardly Ever         216      411     627
##                              247      380
## ------------------------------------------------
## Some Lessons                 669     1174    1843
```

```
##                              726     1118
## ------------------------------------------------
## Total                       3223     4964    8187
## ================================================
##
## Statistics for All Table Factors
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 = 26.77404      d.f. = 3      p = 6.57e-06
##
##         Minimum expected frequency: 246.8329
```

Tablas resumen:

```
##                           CCAA       Sex     Repeated  Homework    Teach_int
## p-value (chi^2) 6.607716e-69 0.7327528 8.825868e-46 0.4182197 6.565762e-06



##                                   CCAA   Sex Repeated Homework Teach_int
## association with TYPE variable:  TRUE FALSE     TRUE    FALSE      TRUE
```

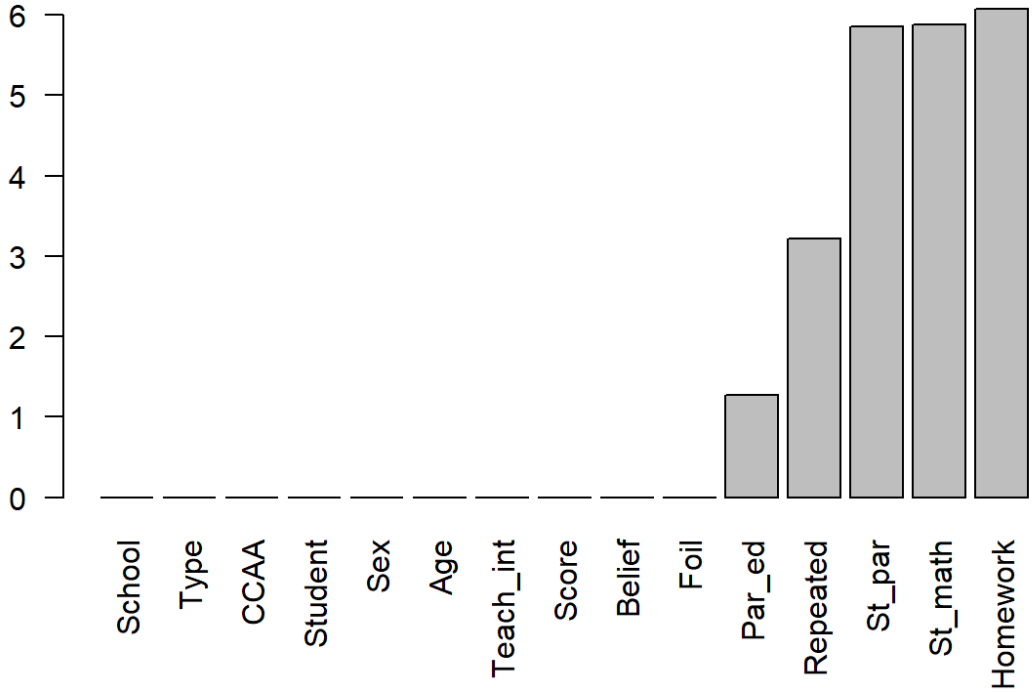
Preprocessing

Formato de datos

Por lo general, no hubo que modificar el formato de los datos debido a que en la propia web de PISA ya nos proveía un archivo de sintaxis de importación (para SPSS y SAS). Los cambios de formato que tuvimos que realizar fue añadir el nivel other a CCAA y recodificar la variable Subnatio en Type (Público/Privado).

Tratamiento de valores faltantes (missings)

5 de las variables de las 15 de nuestra base de datos sobre los resultados del examen PISA de 2012 presentan valores faltantes, que aparecen en 2 variables categóricas (Repetición de curso en secundaría y dedicación a los deberes extraescolar) y 3 variables númericas (Minutos semanales dedicados a estudiar matemáticas, Estudio con al menos uno de los padres y máximos años de educación alcanzados por alguno de los padres).



Estos missings provienen de la estructura de la prueba PISA en sí, que estaba estructurada por un cuestionario común, seguido de una combinación de 2 de 3 cuestionarios extras, y luego la prueba en sí, que constaba de 3 partes, 1 por disciplina a examinar, habiendo por cada una más de 10 versiones distintas.

Las 5 variables en cuestión formaban parte de los cuestionarios extras, y confiando en el proceso de aleatorización en la entrega de estos, consideramos justificable realizar una imputación valores para los valores faltantes en las variables númericas.

Variables categóricas

En las variables categóricas referidas a repetición de curso y dedicación a los deberes, decidimos añadir los valores faltantes como categoría ("Unknown")

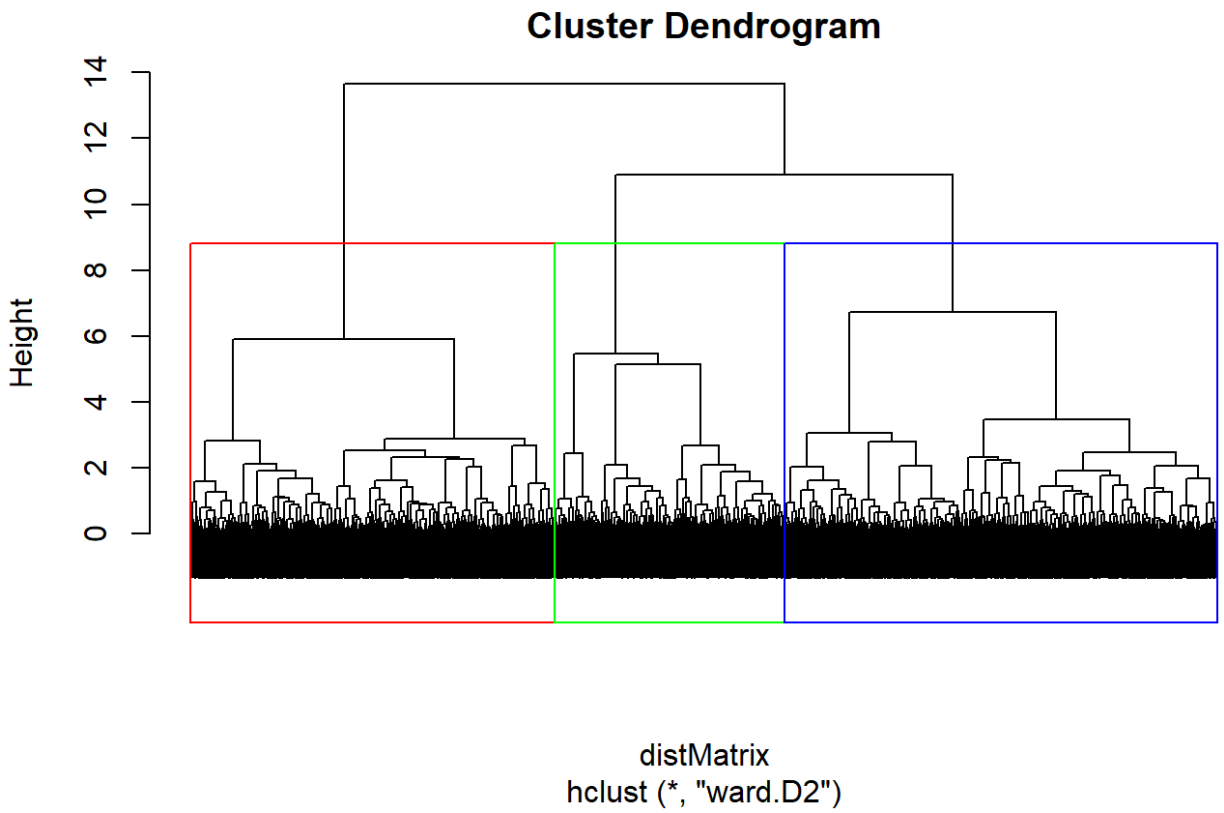
```
        No, never          Yes, once Yes, twice or more              <NA>
             6080               1682                162               263
        No, never          Yes, once Yes, twice or more           Unknown
             6080               1682                162               263
Never or hardly ever Once or twice a month  Once or twice a week      Almost every day
                3183                 1970                 1641                   654
           Every day                 <NA>
                 242                  497
Never or hardly ever Once or twice a month  Once or twice a week      Almost every day
                3183                 1970                 1641                   654
           Every day              Unknown
                 242                  497
```

Variables numéricas

Mixed Intelligent-Multivariate Missing Imputation (MIMMI) method como criterio de imputación

Para imputar los valores faltantes de las variables numéricas, separamos la base de datos entre las filas que contienen algún missing y las que no, y hacemos un clustering de datos mixtos, a través de generar una matriz de disimilitud con la métrica de Gower y utilizar el método de Ward.

Realizamos el dendograma, y en base a las distancias realizamos un corte en 3 grupos y comprobaremos la validez de nuestra decisión en base al criterio de la silueta (que coincide con los 3 grupos). Los clusters nos quedan de tamaños 3450, 2900 y 1800.

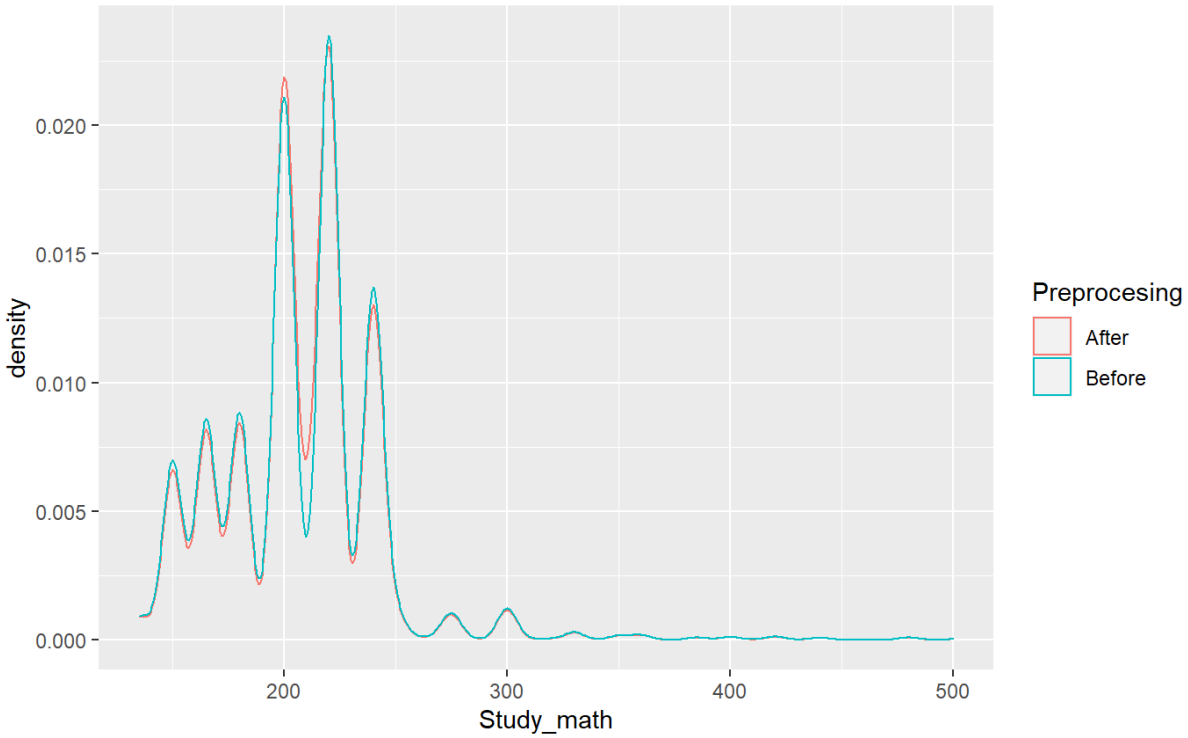


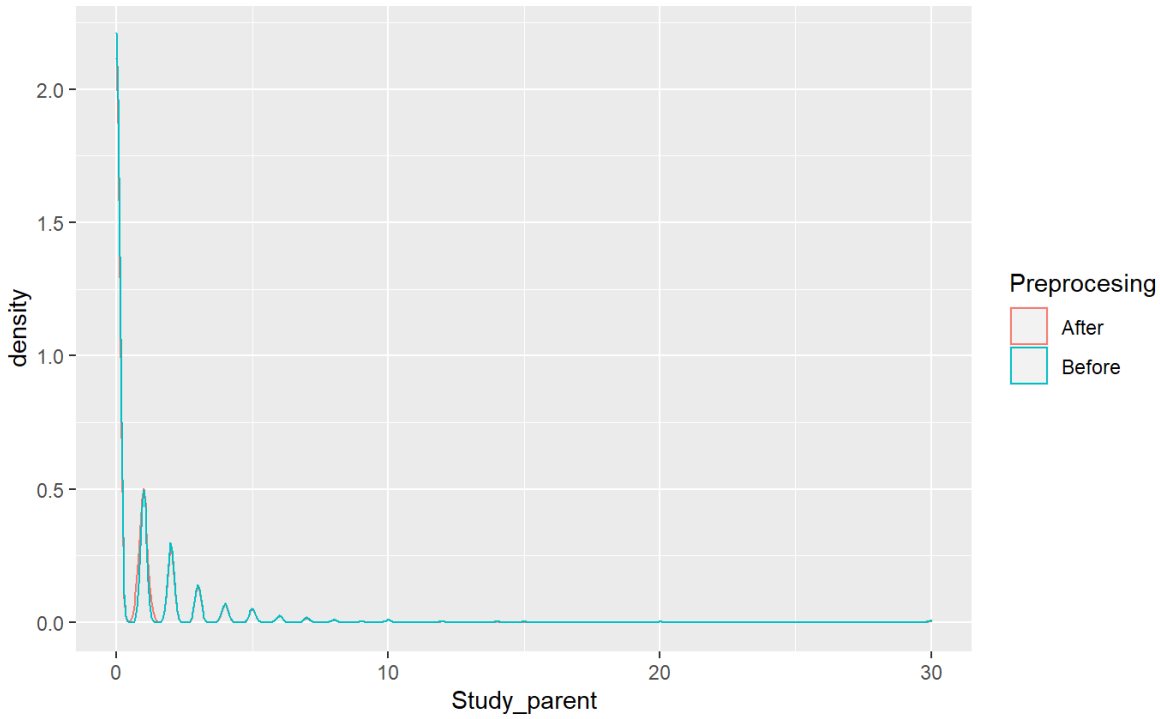

```
 Only frey, mcclain, cindex, sihouette and dunn can
be computed. To compute the other indices, data
matrix is needed
$All.index
     2      3      4      5      6      7
0.1146 0.1181 0.0884 0.0810 0.0913 0.0953
     8      9     10     11     12     13
0.0921 0.0957 0.0800 0.0642 0.0631 0.0590
    14     15
0.0559 0.0490

$Best.nc
Number_clusters     Value_Index
         3.0000          0.1181
```

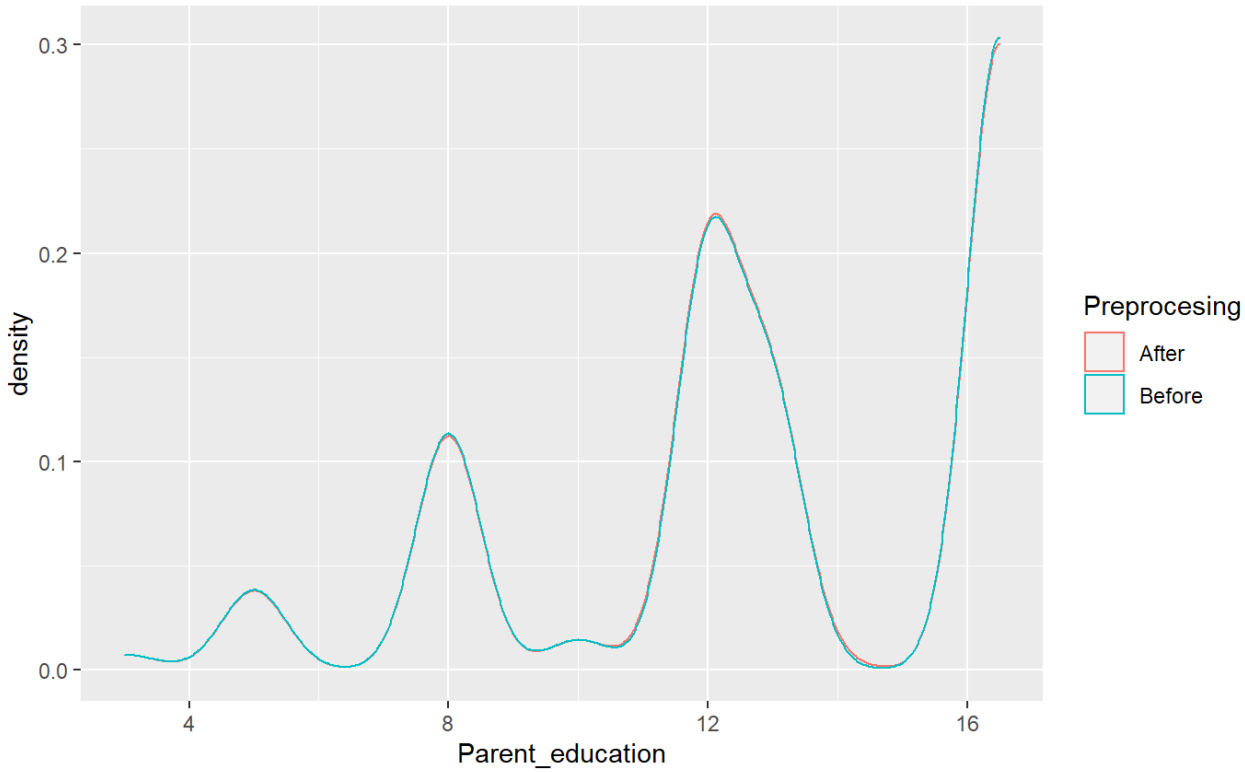
```
c2
   1    2    3
3453 2899 1835
```

Una vez tenemos asignado un grupo para cada individuo, imputaremos en los missings de cada variable la media de su grupo en esa variable.

Para verificar la calidad de la imputación en las 3 variables numéricas, comparamos sus densidades previas al preprocesing y después, y observamos una gran similitud en los 3 casos.






0.3
0.2
0.1
0.0
density
4
8
12
16
Parent_education
Preprocesing
After
Before

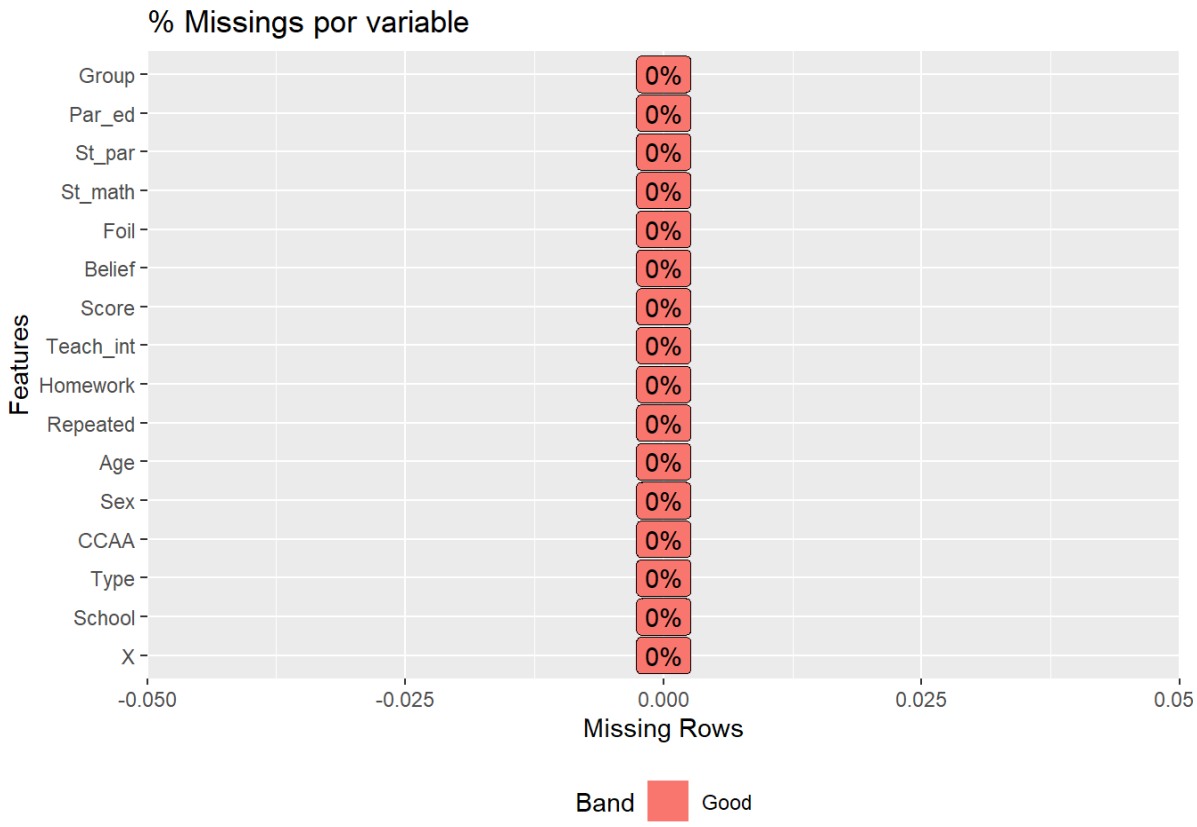
Descriptiva Univariante después del Preprocessing

Una vez hecho el preprocessing de nuestra base de datos, volvemos a hacer una descriptiva de nuestras variables para comprobar si muestran cambios destacables.

Missings

Por lo que respecta al porcentaje de missings podemos ver que las variables han sido correctamente imputadas y que no presentan ningún dato faltante. Observamos también que en lugar de la variable Student, tenemos ahora la variable Group asociada a la imputación usando el método MIMMI.

Figura: Missings de las variables



Variables Numéricas

Estadísticos principales

En cuanto a las variables numéricas se han producido ligeros cambios en los valores de aquellas variables que tenían missings y que han sido imputados. Aún así no hay ningún cambio muy destacado.

Tabla 1: Estadísticos variables numéricas

Variables	n_missing	complete_rate	mean	sd
Age	0	1	15.87	0.29
Score	0	1	498.14	84.65
Belief	0	1	0.53	1.21
Foil	0	1	-0.52	1.17
St_math	0	1	207.03	36.50
St_par	0	1	0.89	2.15
Par_ed	0	1	12.84	3.53

Tabla 2: Estadísticos variables numéricas

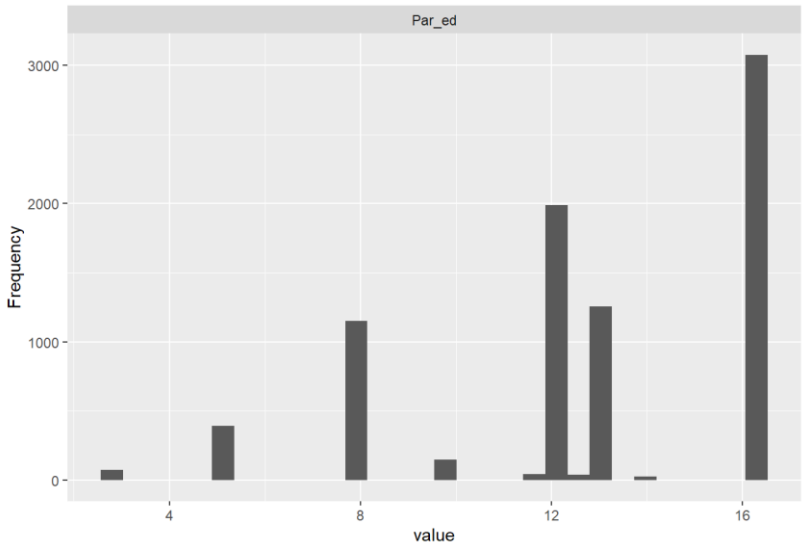
Variables	mín	p25	p50	p75	máx
Age	15.33	15.58	15.92	16.08	16.33
Score	130.06	441.04	500.40	558.86	763.13
Belief	-5.25	-0.18	0.48	1.15	6.00
Foil	-2.31	-1.47	-0.80	0.30	4.24
St_math	135.00	180.00	201.20	220.00	500.00
St_par	0.00	0.00	0.00	1.00	30.00
Par_ed	3.00	12.00	13.00	16.50	16.50

Histogramas

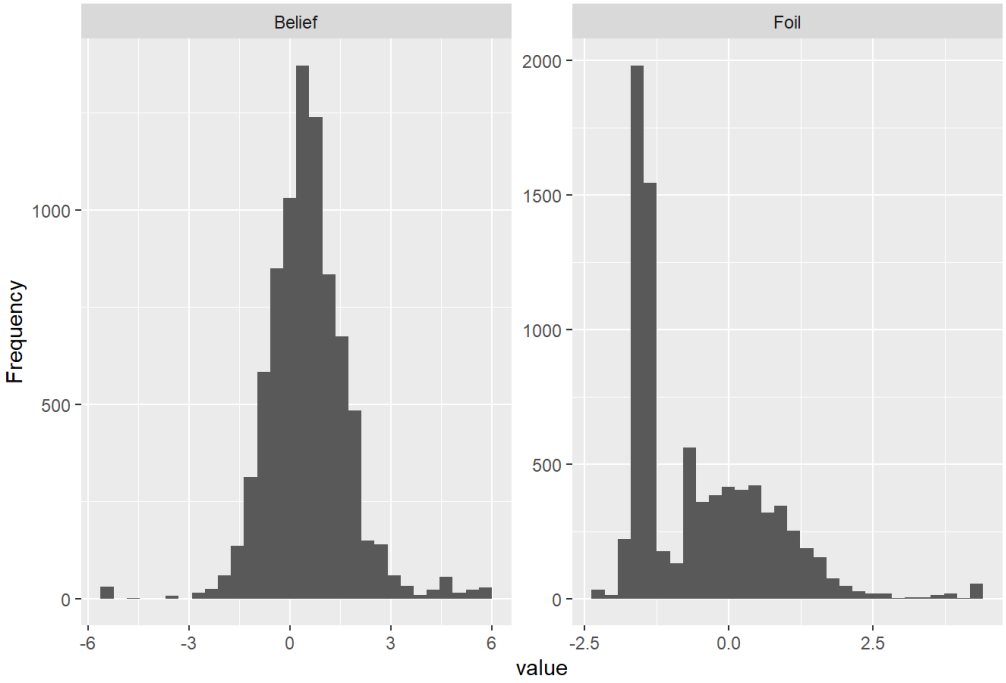
Por lo que respecta a los histogramas comentaremos aquellos que hayan sufrido algún cambio.

Es el caso del histograma de Belief, ahora podemos apreciar mejor una tendencia a seguir a una distribución del tipo normal. Por lo que respecta al resto de histogramas no se aprecia ningún cambio relevante.

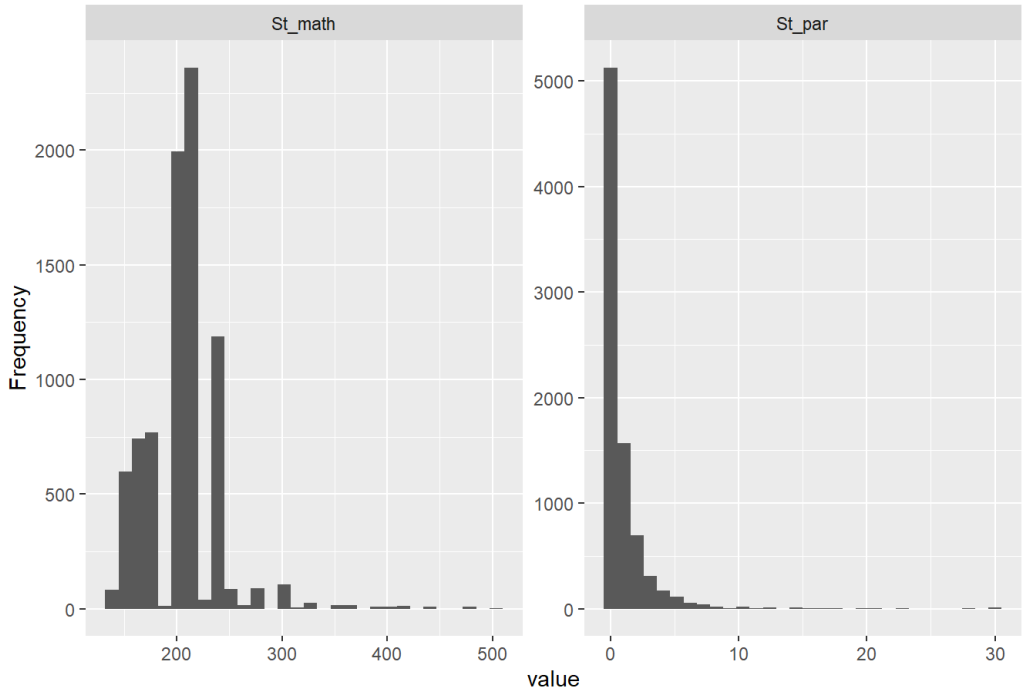
Histogramas : Par_ed



Histogramas: Belief y Foil



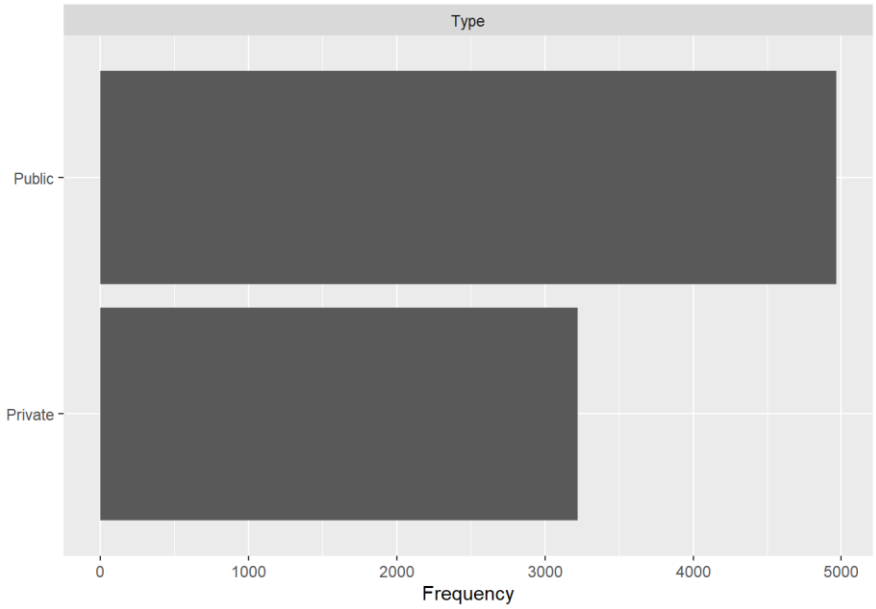
Histogramas: St_math y St_par

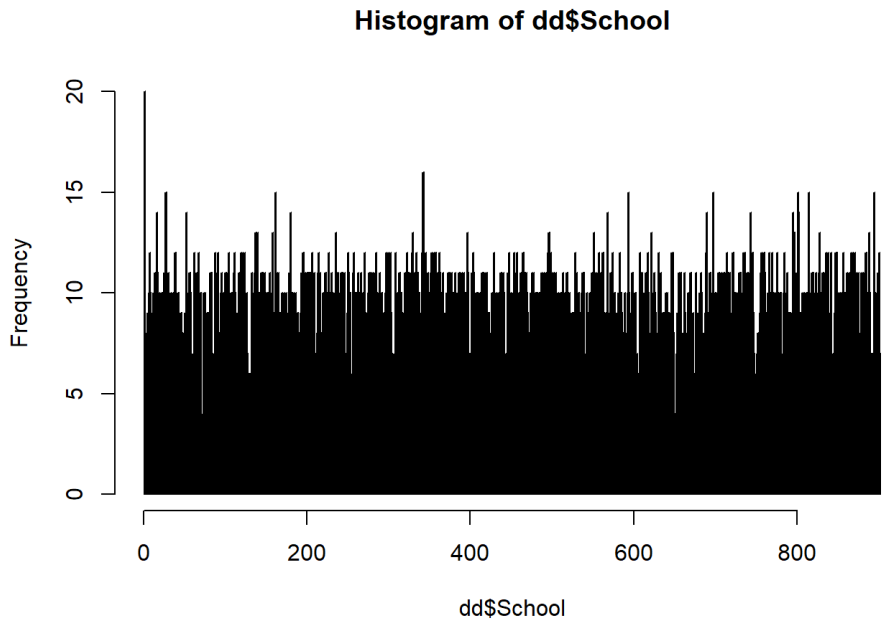


Variables Categóricas

Por lo que respecta a variables categóricas también comentaremos las tablas de frecuencia dónde se haya producido algún cambio.

Tabla de Frecuencias : Type





Puesto que la variable School, la consideramos ahora como un factor, la situamos junto con las variables categóricas. En un gráfico probablemente más claro que el que podíamos observar en la descriptiva antes del preprocessing.

Tabla de Frecuencias : CCAA y Sex

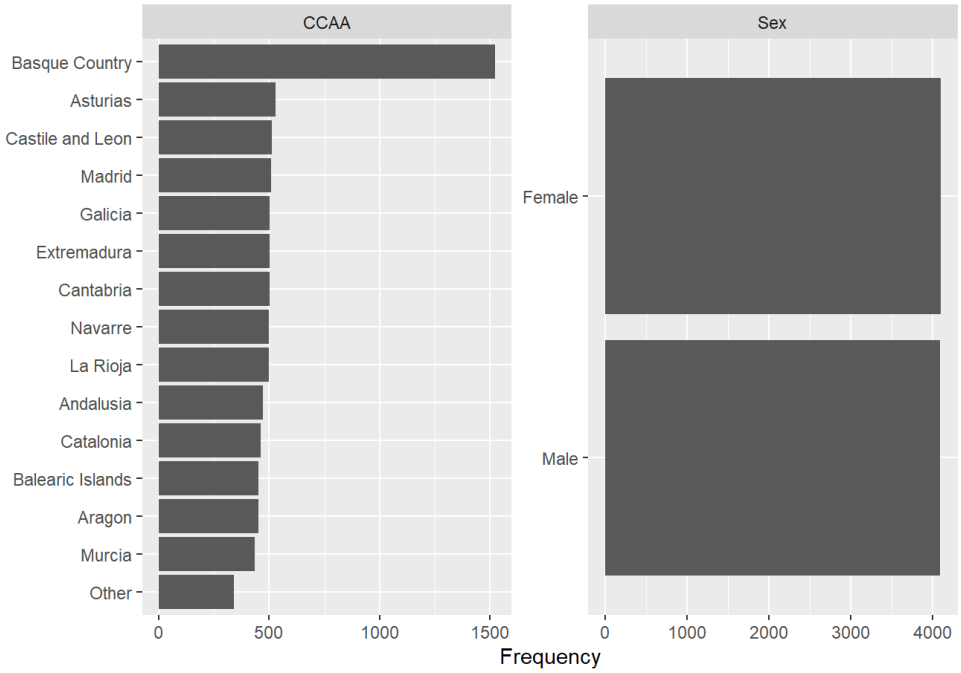
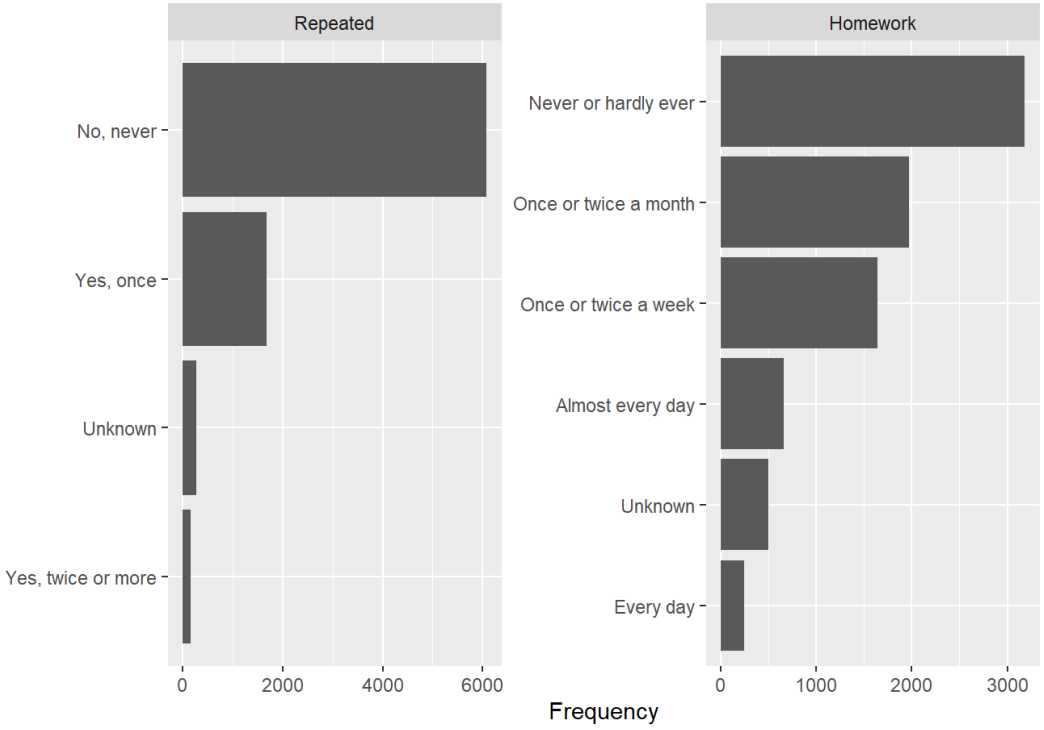
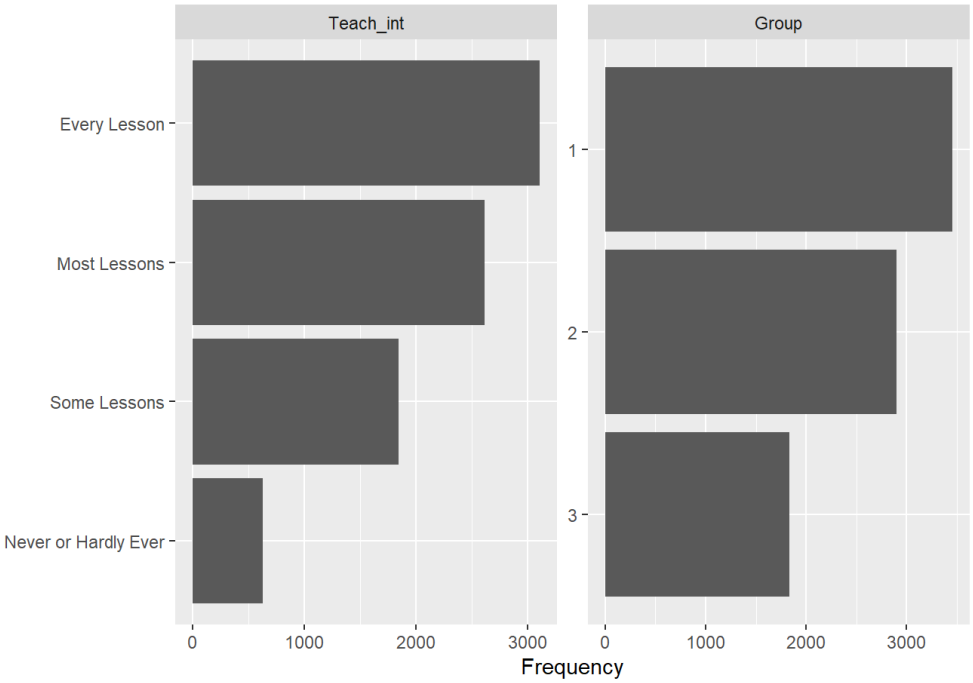


Tabla de Frecuencias : Repeated y Homework



Los cambios que observamos són respecto al hecho de añadir una nueva categoría "Unknown", producto de la imputación de los datos.

Tabla de Frecuencias : Type y Teach_int



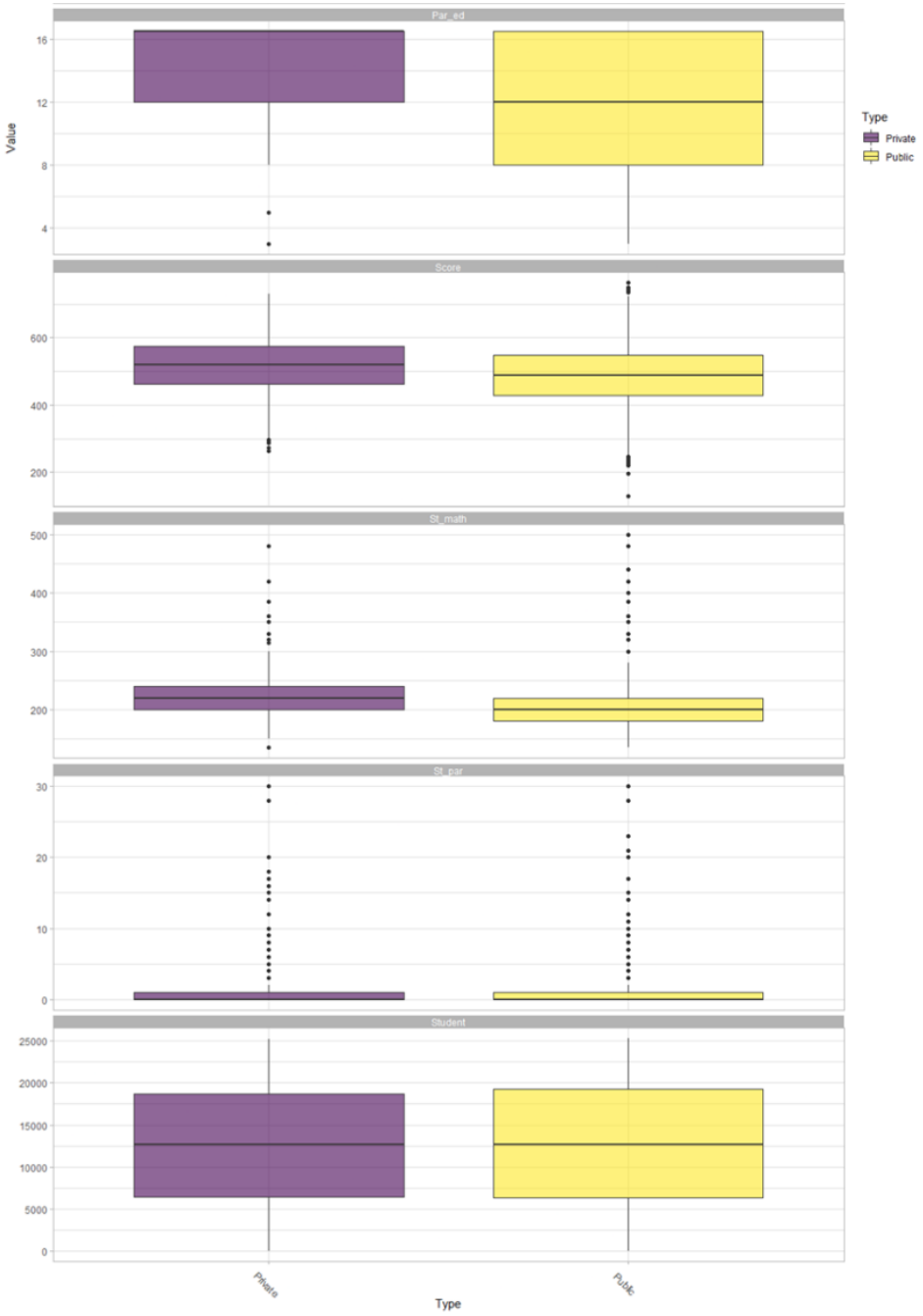
Finalmente vemos que en la nueva variable Group tiene 3 categorías, de las cuales el grupo 1 es la predominante. No són grupos totalmente equitativos.

Descriptiva Bivariante después del Preprocessing

A continuación realizaremos un análisis descriptivo bivariante entre las variables numéricas de nuestra base de datos con la variable respuesta, Type. Miraremos si existen diferencias entre los diferentes tipos de escuela por cada variable numérica.

Por estos gráficos el amarillo representa la escuela pública y el lila la escuela privada.

Type:


Par_ed
Score
St_math
St_par
Student
Value
Type
Private
Public
Private
Public
Type

A continuación comentaremos las diferencias más significativas de los distintos tipos de escuela con las variables numéricas.

ST_MATH: Por la variable St_math vemos que el rango Intercuartílico de los datos del tipo de escuela pública está ligeramente inferior al de la escuela privada. Concretamente el rango IQR inferior de la escuela privada comienza donde se sitúa el punto medio, o promedio, de la pública. Por eso vemos que los datos privados estarían alrededor de 225 minutos mientras que las públicas estarían alrededor de 200 minutos. Recordemos que la variable St_math son los minutos que el estudiante suele dedicar semanalmente al estudio a las matemáticas.

PAR_ED: En el caso de años máximos de educación de los padres vemos que al igual que en el caso anterior los estudiantes de la escuela privada tienen unos padres que han logrado estudiar un máximo de años superior. Por ejemplo, al igual que antes, el punto del IQR inferior del tipo privado se encuentra donde se sitúa la media del público. También podemos ver que respecto a la media de los dos hay cuatro años de diferencia, los padres de los estudiantes de la escuela pública solo llegan a los 12 años mientras que por la privada llegan al máximo, 16.

Por el resto de variables nos encontramos las mismas situaciones, los rangos intercuartilicos al igual que las medias se encuentran ligeramente superior de forma general.

Tabla 4: Estadísticos variables numéricas

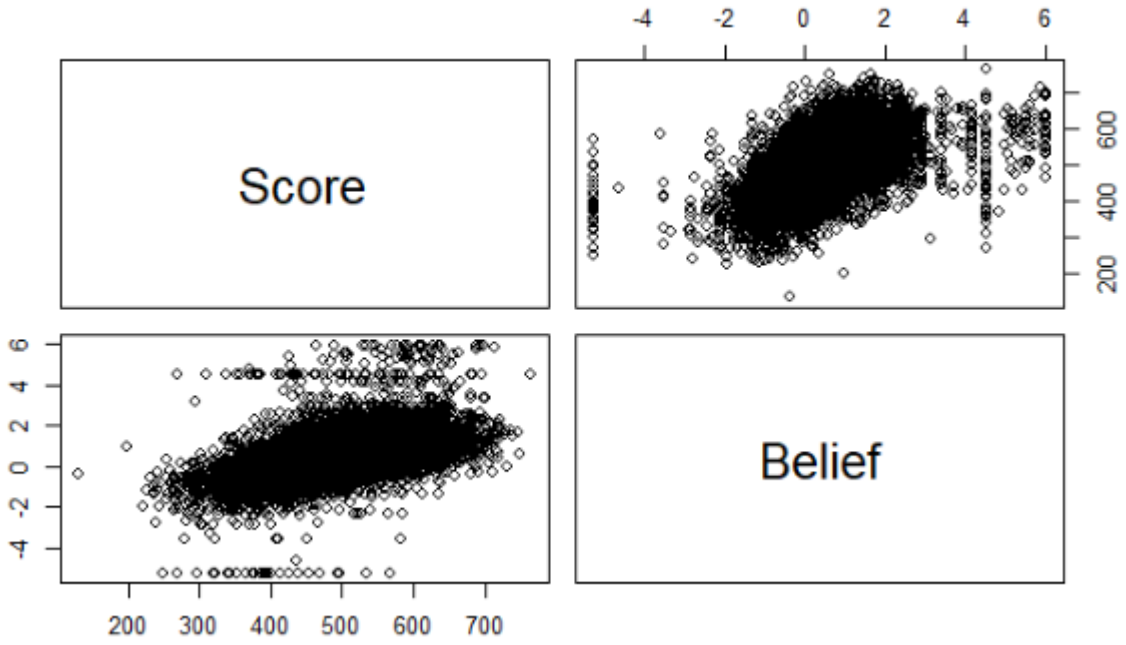
variable	Type	N	Mean	SD	Min	Max	Median	Q1	Q3	p.value
Student	Private	3,223	12,608.186	7,368.436	37.000	25,250.000	12,684.000	6,409.500	18,728.500	0.4027
Student	Public	4,964	12,746.655	7,279.161	3.000	25,309.000	12,704.500	6,339.750	19,243.750	
Age	Private	3,223	15.874	0.288	15.330	16.330	15.920	15.670	16.080	0.2414
Age	Public	4,964	15.867	0.290	15.330	16.330	15.920	15.580	16.080	
Score	Private	3,223	516.690	78.957	263.522	730.885	520.026	462.852	573.773	< 0.0001
Score	Public	4,964	486.100	86.050	130.059	763.133	487.545	427.313	547.620	
Belief	Private	3,223	0.786	1.153	-5.250	5.998	0.709	0.129	1.360	< 0.0001
Belief	Public	4,964	0.367	1.224	-5.250	5.998	0.331	-0.386	1.002	
Foil	Private	3,223	-0.561	1.166	-2.312	4.244	-1.218	-1.470	0.278	0.0189
Foil	Public	4,964	-0.499	1.165	-2.312	4.244	-0.715	-1.487	0.309	
St_math	Private	3,223	211.239	33.269	135.000	480.000	220.000	200.000	240.000	< 0.0001
St_math	Public	4,964	204.303	38.212	135.000	500.000	200.000	180.000	220.000	
St_par	Private	3,223	0.828	2.035	0.000	30.000	0.000	0.000	1.000	0.0509
St_par	Public	4,964	0.923	2.226	0.000	30.000	0.000	0.000	1.000	
Par_ed	Private	3,223	13.846	3.112	3.000	16.500	16.500	12.000	16.500	< 0.0001
Par_ed	Public	4,964	12.185	3.629	3.000	16.500	12.000	8.000	16.500	
Group	Private	3,223	2.111	0.347	1.000	3.000	2.000	2.000	2.000	< 0.0001
Group	Public	4,964	1.602	0.906	1.000	3.000	1.000	1.000	3.000	

Descriptiva Bivariante Numérica-Numérica

Por las variables numéricas calculamos las correlaciones por parejas. Miramos las correlaciones máximas para así encontrar las variables con más correlación.

	Student	Age	Score	Belief	Foil	St_math	St_par	Par_ed	Group
Student	1.000000000	0.002363755	-0.007461227	0.009563617	-0.01942877	-0.026927404	0.005951831	-0.01521280	-0.002968728
Age	0.002363755	1.000000000	0.043825445	0.042947892	0.03975067	-0.006528083	-0.026071619	0.01540523	-0.033065846
Score	-0.007461227	0.043825445	1.000000000	0.514223121	-0.01609374	-0.068920921	-0.133938691	0.28908431	-0.366723830
Belief	0.009563617	0.042947892	0.514223121	1.000000000	0.12601534	-0.046461470	-0.035769463	0.21721323	-0.296313990
Foil	-0.019428768	0.039750666	-0.016093741	0.126015338	1.00000000	0.016098066	0.050155907	0.05934535	-0.039591851
St_math	-0.026927404	-0.006528083	-0.068920921	-0.046461470	0.01609807	1.000000000	0.004822718	-0.03055982	0.127686748
St_par	0.005951831	-0.026071619	-0.133938691	-0.035769463	0.05015591	0.004822718	1.000000000	0.02903876	0.069121440
Par_ed	-0.015212805	0.015405227	0.289084309	0.217213233	0.05934535	-0.030559824	0.029038763	1.00000000	-0.076040349
Group	-0.002968728	-0.033065846	-0.366723830	-0.296313990	-0.03959185	0.127686748	0.069121440	-0.07604035	1.000000000

La correlación maxima corresponde a Score i Belief aunque no destaca mucho ya que con dificultades passa el valor de 0.5. También vemos que al contrario de antes han desaparecido todos los NA’s que ocupaban media tabla de las correlaciones. A continuación vemos el gráfico de dispersión de las variables con más correlación.

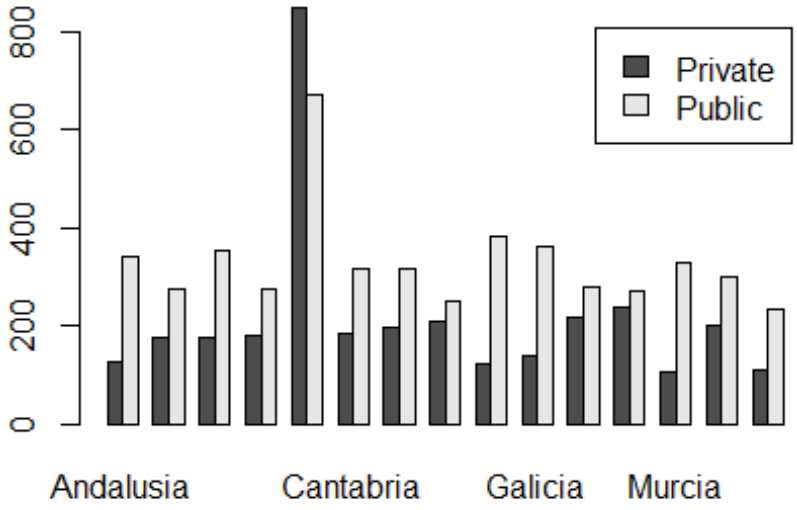


Descriptiva Bivariante Categórica-Categórica

Para analizar si existe asociación entre las variables categóricas de nuestra base de datos y Type, la variable respuesta, haremos tablas de contingencia (con sus valores esperados) y el

test chi cuadrado con un nivel de significación de 0.05. Las variables explicativas aquí analizadas son: CCAA, Sex, Repeated, Homework, y Teach_int.

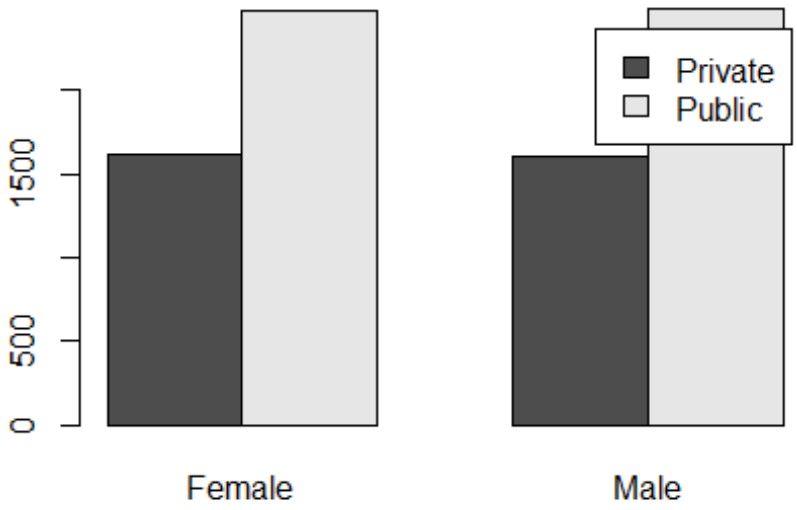
CCAA




```
##    Cell Contents
## |-------------------------|
## |                   Count |
## |         Expected Values |
## |-------------------------|
##
## ============================================
##                     ddcat$Type
## ddcat[[k]]          Private   Public   Total
## --------------------------------------------
## Andalusia               128      343     471
##                         185      286
## --------------------------------------------
## Aragon                  177      274     451
##                         178      274
## --------------------------------------------
## Asturias                174      355     529
##                         208      321
## --------------------------------------------
## Balearic Islands        178      275     453
##                         178      275
## --------------------------------------------
## Basque Country          849      674    1523
##                         600      923
```

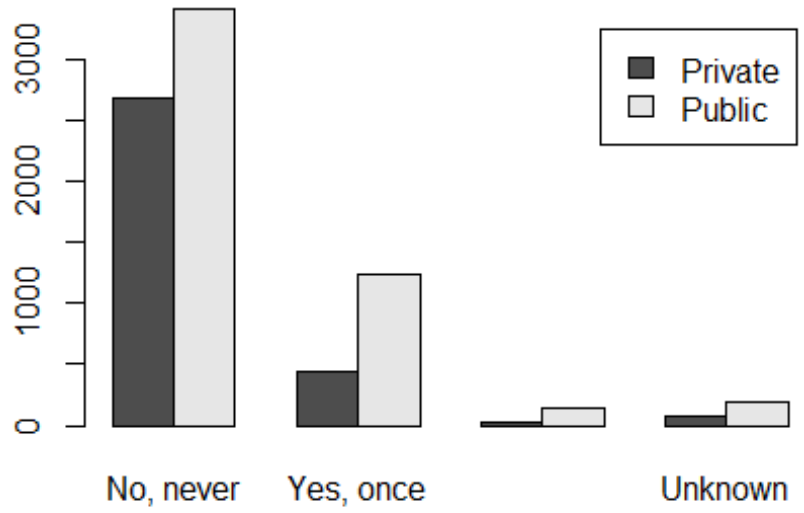
```
## ---------------------------------------------
## Cantabria               186      316     502
##                         198      304
## ---------------------------------------------
## Castile and Leon        197      316     513
##                         202      311
## ---------------------------------------------
## Catalonia               210      252     462
##                         182      280
## ---------------------------------------------
## Extremadura             121      381     502
##                         198      304
## ---------------------------------------------
## Galicia                 138      364     502
##                         198      304
## ---------------------------------------------
## La Rioja                217      280     497
##                         196      301
## ---------------------------------------------
## Madrid                  237      271     508
##                         200      308
## ---------------------------------------------
## Murcia                  104      329     433
##                         170      262
## ---------------------------------------------
## Navarre                 199      301     500
##                         197      303
## ---------------------------------------------
## Other                   108      233     341
##                         134      207
## ---------------------------------------------
## Total                  3223     4964    8187
## =============================================
##
## Statistics for All Table Factors
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 = 363.314      d.f. = 14      p <2e-16
##
##         Minimum expected frequency: 134.2425
```

SEX



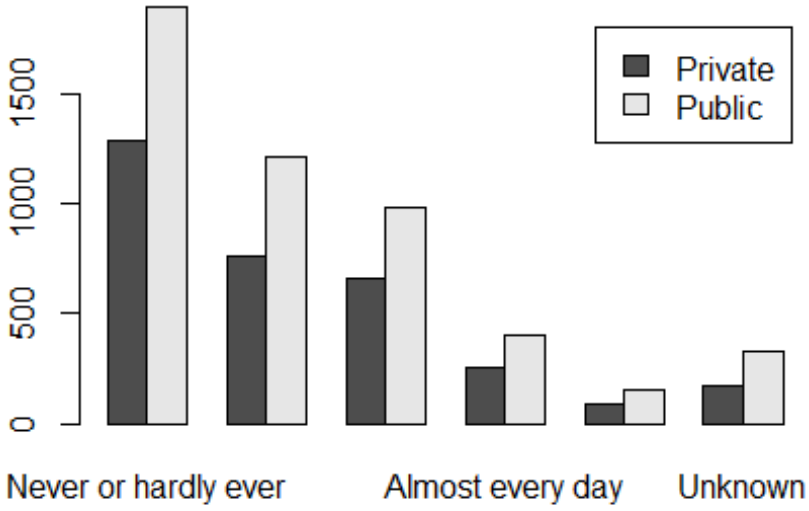

```
##    Cell Contents
## |-------------------------|
## |                   Count |
## |         Expected Values |
## |-------------------------|
##
## ====================================
##               ddcat[[k]]
## ddcat$Type    Female   Male   Total
## ------------------------------------
## Private         1622   1601    3223
##                 1614   1608
## ------------------------------------
## Public          2479   2485    4964
##                 2486   2478
## ------------------------------------
## Total           4101   4086    8187
## ====================================
##
## Statistics for All Table Factors
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 = 0.1165991      d.f. = 1      p = 0.733
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 = 0.101662      d.f. = 1      p = 0.75
##         Minimum expected frequency: 1608.547
```

REPEATED



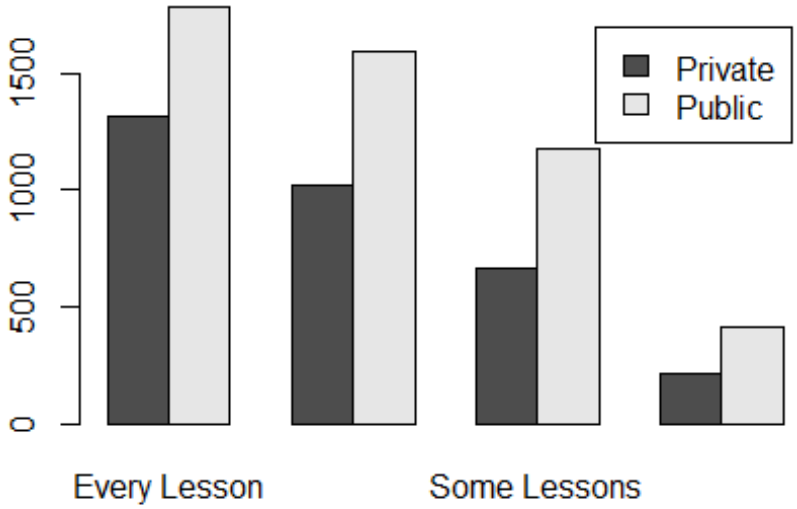

```
##    Cell Contents
## |-------------------------|
## |                   Count |
## |         Expected Values |
## |-------------------------|
##
## ============================================================================
##               ddcat[[k]]
## ddcat$Type    No, never   Yes, once   Yes, twice or more   Unknown   Total
## ----------------------------------------------------------------------------
## Private            2675         442                   28        78    3223
##                    2394         662                   64       104
## ----------------------------------------------------------------------------
## Public             3405        1240                  134       185    4964
##                    3686        1020                   98       160
## ----------------------------------------------------------------------------
## Total              6080        1682                  162       263    8187
## ============================================================================
##
## Statistics for All Table Factors
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 = 218.8013      d.f. = 3      p <2e-16
##
##         Minimum expected frequency: 63.77501
```

HOMEWORK




```
##    Cell Contents
## |-------------------------|
## |                   Count |
## |         Expected Values |
## |-------------------------|
##
## ============================================================================
##            ddcat[[k]]
## ddct$Ty    N o h e   O o t a   O o t a   Alm e d   Evry dy   Unknown   Total
## ----------------------------------------------------------------------------
## Private       1290       761       658       254        86       174    3223
##               1253       776       646       258        95       196
## ----------------------------------------------------------------------------
## Public        1893      1209       983       400       156       323    4964
##               1930      1194       995       396       147       301
## ----------------------------------------------------------------------------
## Total         3183      1970      1641       654       242       497    8187
## ============================================================================
##
## Statistics for All Table Factors
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 = 8.128985      d.f. = 5      p = 0.149
##
##         Minimum expected frequency: 95.26884
```

TECH_INT




```
##    Cell Contents
## |-------------------------|
## |                   Count |
## |         Expected Values |
## |-------------------------|
##
## ================================================
##                          ddcat$Type
## ddcat[[k]]               Private   Public   Total
## ------------------------------------------------
## Every Lesson                1320     1784    3104
##                             1222     1882
## ------------------------------------------------
## Most Lessons                1018     1595    2613
##                             1029     1584
## ------------------------------------------------
## Some Lessons                 669     1174    1843
##                              726     1118
## ------------------------------------------------
## Never or Hardly Ever         216      411     627
##                              247      380
## ------------------------------------------------
## Total                       3223     4964    8187
## ================================================
##
## Statistics for All Table Factors
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
```

```
## Chi^2 = 26.77404      d.f. = 3      p = 6.57e-06
##
##         Minimum expected frequency: 246.8329
```

Interpretación y resumen

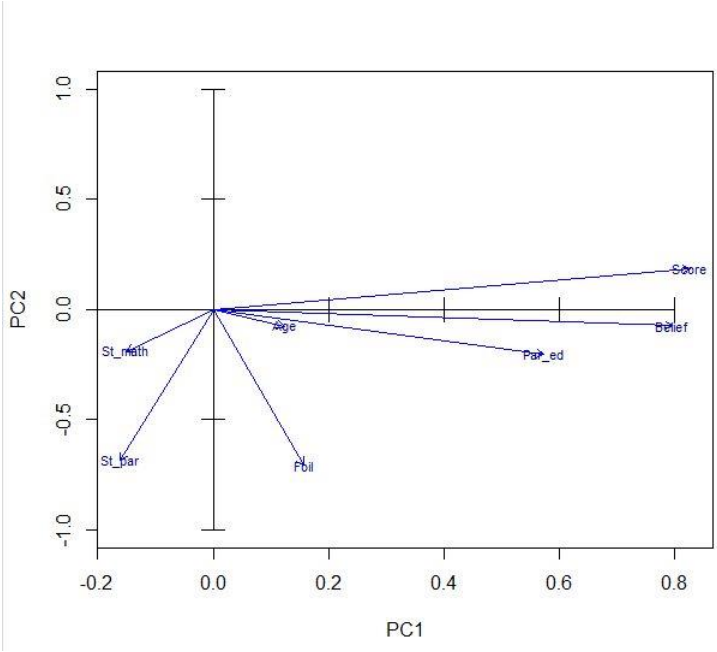
Resumimos lo visto anteriormente en las siguientes tablas:

```
##                           CCAA       Sex    Repeated  Homework    Teach_int
## p-value (chi^2) 6.607716e-69 0.7327528 3.64623e-47 0.1492688 6.565762e-06
```

```
##                                   CCAA   Sex Repeated Homework Teach_int
## association with TYPE variable:  TRUE FALSE     TRUE    FALSE      TRUE
```

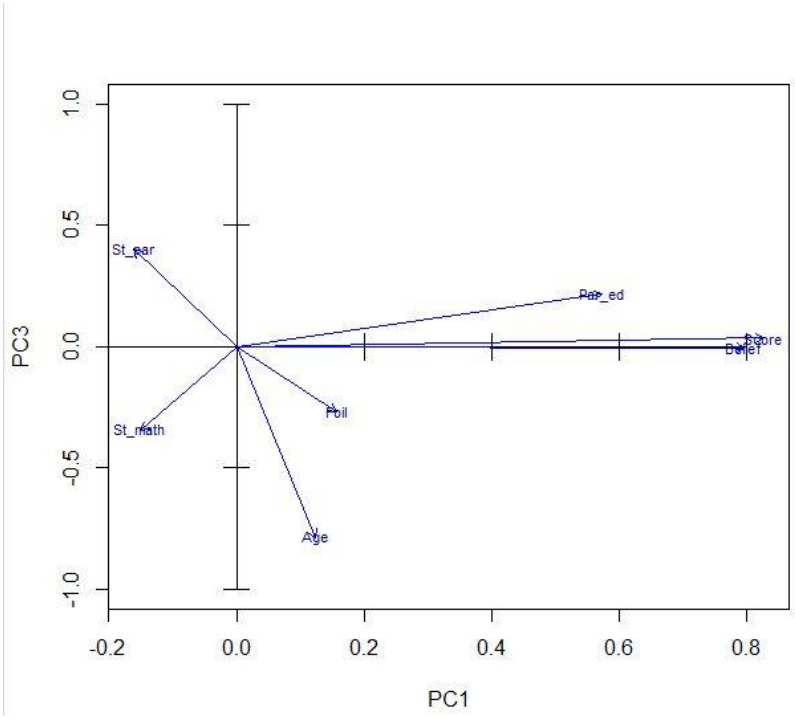
Con un nivel de significación de 0.05, podemos decir que existe asociación entre el tipo de escuela y la comunidad autónoma, el tipo de escuela y el número de veces que repite, y el tipo de escuela y el interés del profesor. En cambio, no existe asociación par el sexo y las veces de deberes que hace fuera de clase por semana respecto el tipo de escuela.

PCA (Principal Component Analysis)

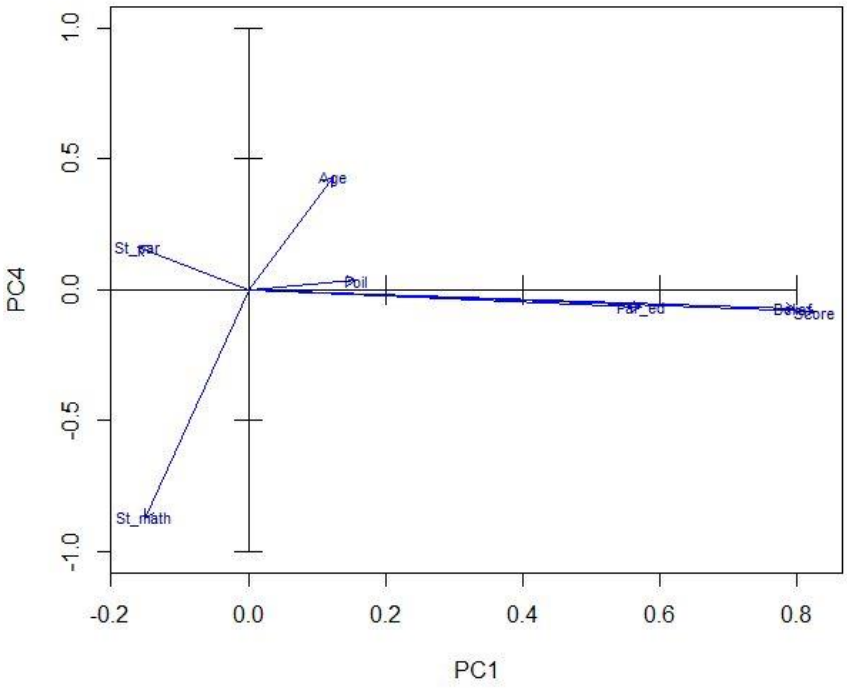


Las variables Belief,Score y Par_ed están correlacionadas positivamente con el PC1, y a la vez están relacionadas entre ellas, Belief y Score son las más importantes para PC1, por lo tanto puede tener que ver con la puntuación (cuanto más creen que saben mayor nota esperada).

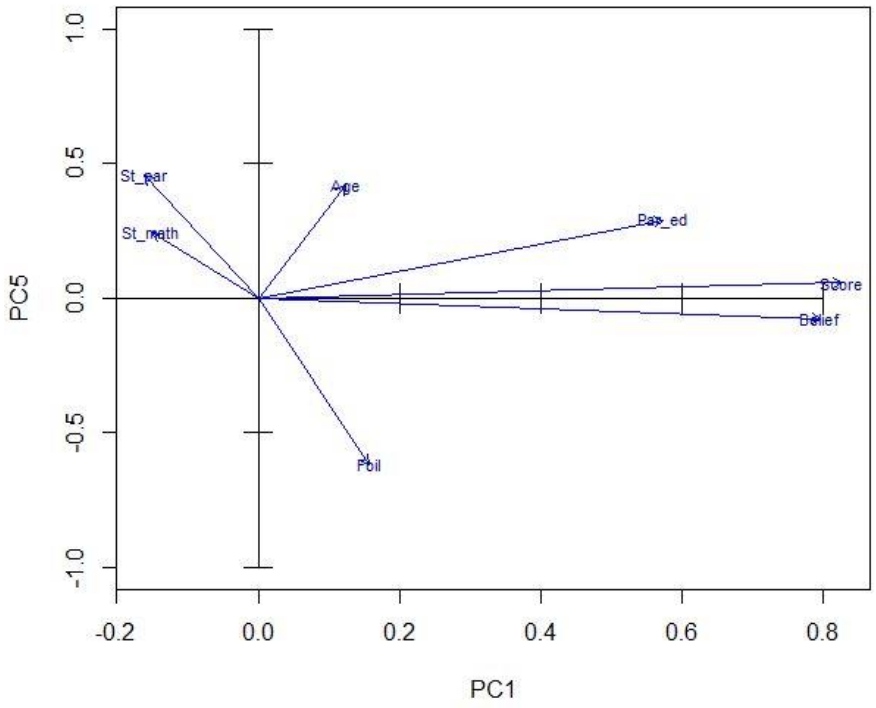
Las variables Foil y St_par están correlacionadas negativamente con la PC2 y también son las más importantes para este plano.



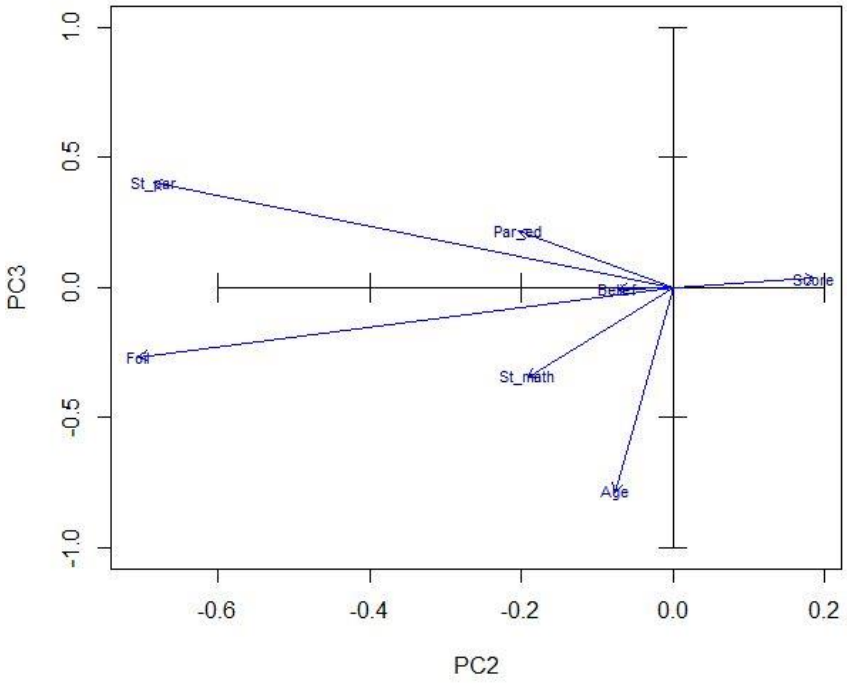
La PC1 está correlacionada con las variables Par_ed, Score y Belief, Score y Belief están muy correlacionadas entre sí y son las importantes para este plano. La PC3 está correlacionada negativamente con la variable Edad.



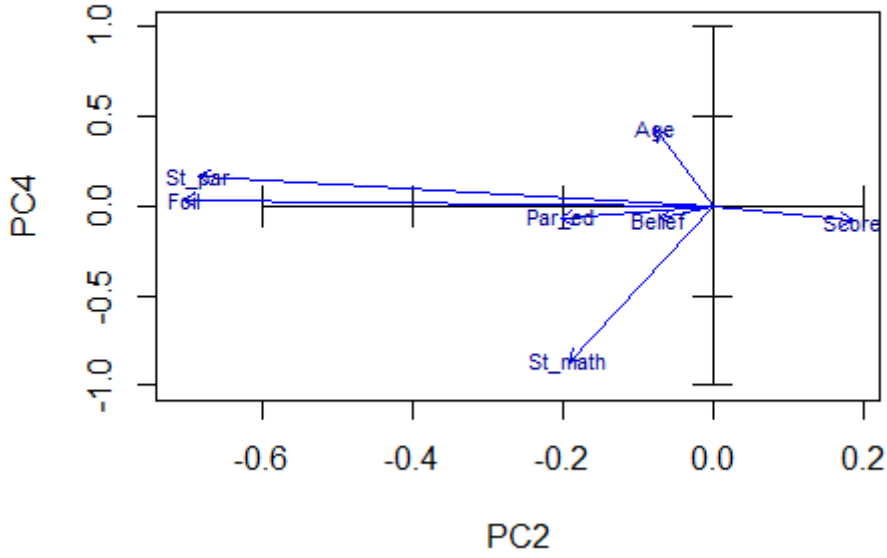
La variable Foil parece estar relacionada positivamente con la PC1, y a la vez relacionada con las variables Par_ed,belief y Score, las cuales están muy correlacionadas entre sí y la variable St_par correlacionada negativamente aunque las más importantes sean Belief y Score.
La PC4 está correlacionada negativamente con la variable St_math siendo esta la más importante.



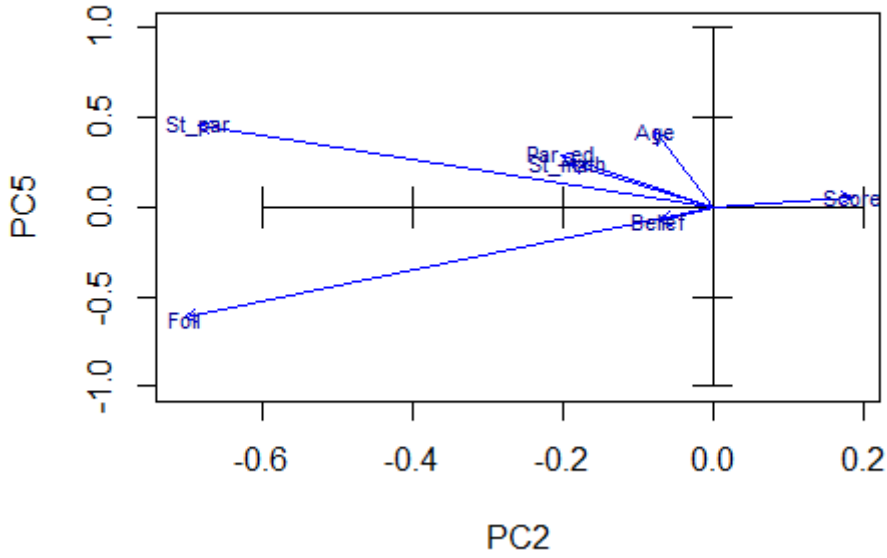
La PC1 está correlacionada con las variables Belief, Score i Par_ed, las variables Belief y Score están correlacionadas entre sí y son las más importantes para este plano. La PC5 está correlacionada con la variable Foil negativamente y es la variable más relevante para el plano.



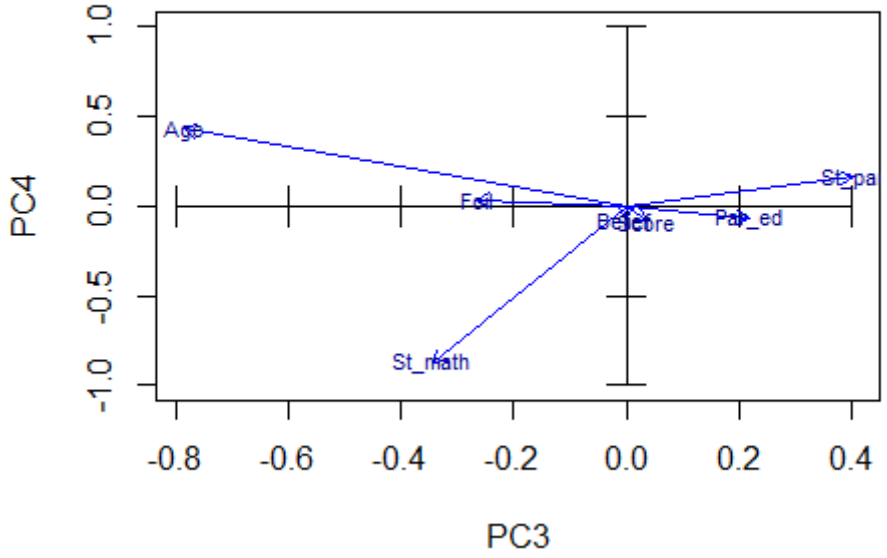
Las variables Belief, St_par y Foil están correlacionadas negativamente con la PC2 y la variable Score positivamente. Las variables Par_ed, St_par y Belief parecen estar correlacionadas entre sí y la variable Foil parece estar correlacionada con las variables Belief, St_par y St_math. Las variables más relevantes para la PC2 son St_par y Foil. La variable Age está correlacionada negativamente con la PC3 y es la variable más relevante para este plano.



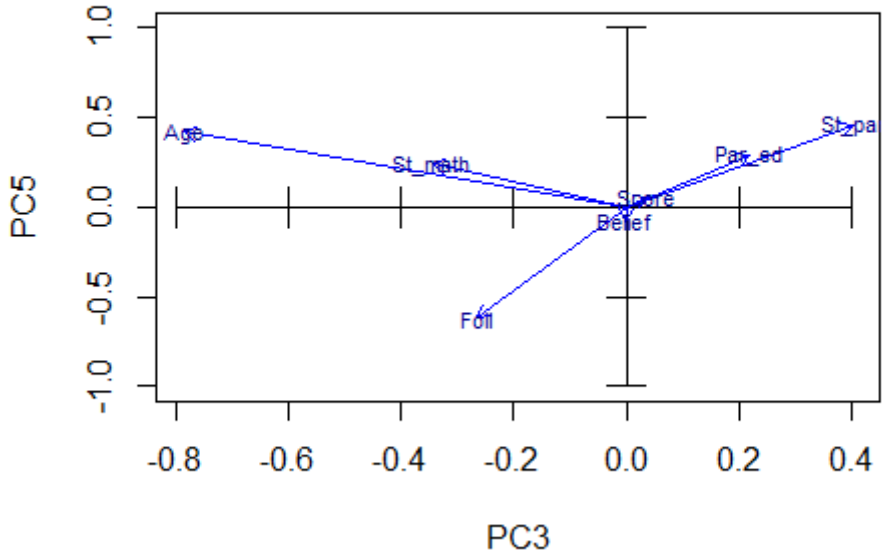
St_par y Foil son las dos variables más importantes para PC2, cuya correlación es negativa. Para PC4, existe una correlación negativa con St_math.



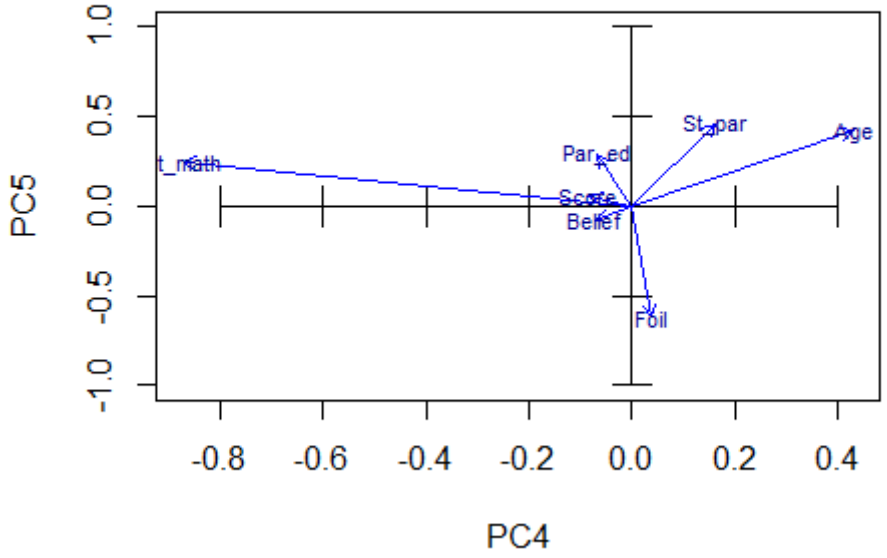
Podemos observar que Age está correlacionada positivamente con PC2. Después, Foil y St_Par son las más relevantes en este plano.



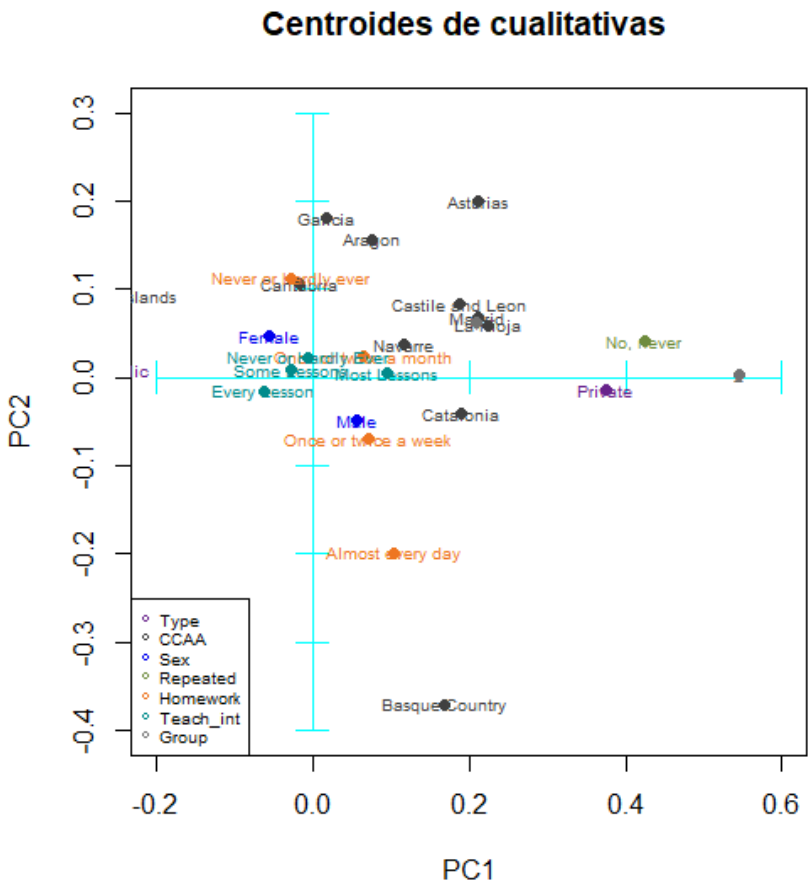
Vemos que Age y St_math son muy relevantes y están correlacionadas negativamente con PC3.



Destacamos las variables Age y St_par, estando negativa y positivamente correlacionadas respectivamente con PC3. PC5 está correlacionada negativamente con Foil y de manera positiva con St_par.

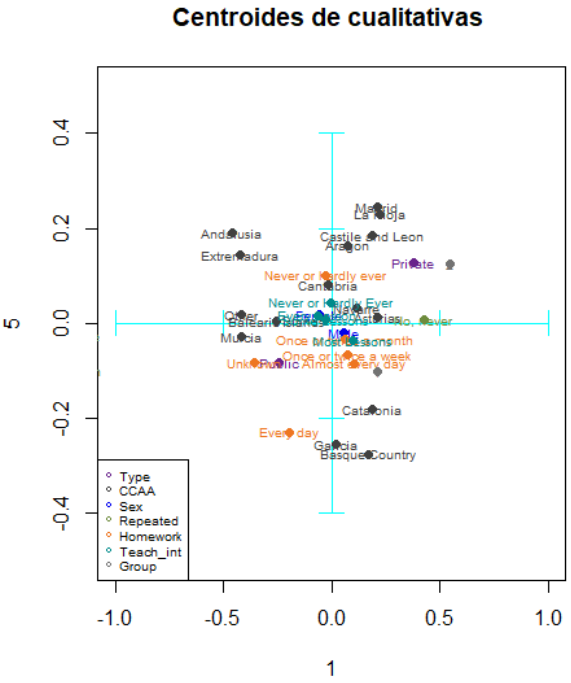


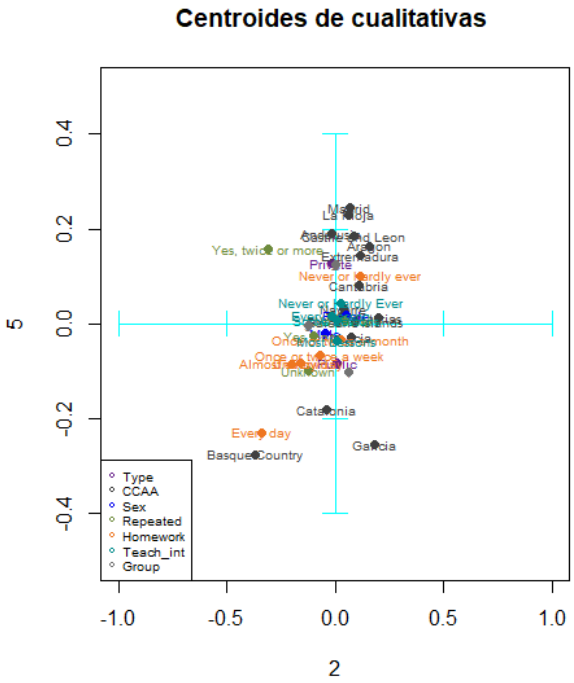
Por último, en cuanto a PC4, está correlacionada negativamente con St_math (siendo la variable más importante) y positivamente con Age. Por otro lado, Foil está correlacionada negativamente con PC5.



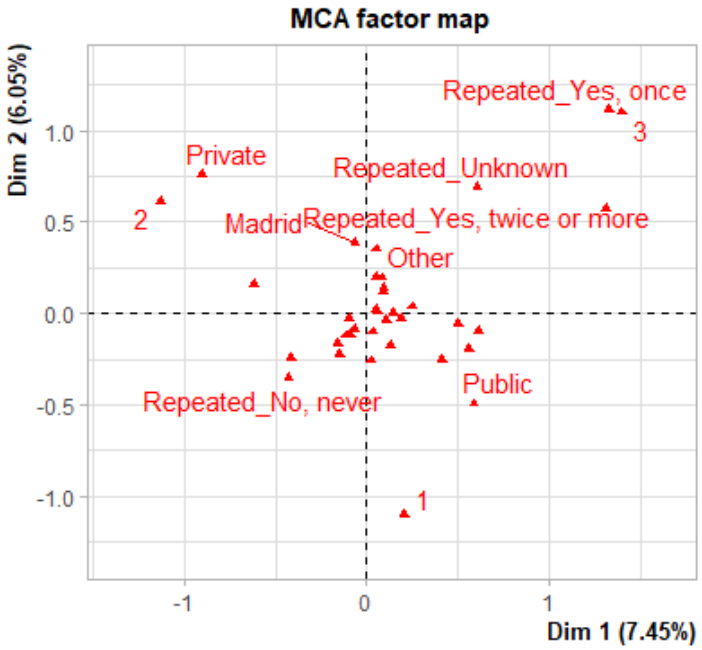
En este gráfico de los dos planos que explican más variabilidad vemos que Cataluña y el País Basco se diferencia de las demás comunidades, aunque el País Basco se diferencia más de todas las demás, que hay diferencias entre géneros, entre los que son de escuela privada y pública (izquierda del gráfico) y entre los que suelen hacer los deberes frecuentemente y los que no.

Resumiendo, podríamos decir que la PC1 puede tener que ver con la puntuación (cuanto más creen que saben mayor nota), la PC2 con el tiempo que dedican a los estudios fuera de clase, la PC3 tiene que ver con la edad, la PC4 con el tiempo que dedican a estudiar las matemáticas y la PC5 tiene algo que ver con las comunidades autónomas (Cataluña, Galicia y País Vasco se diferencian de las demás).

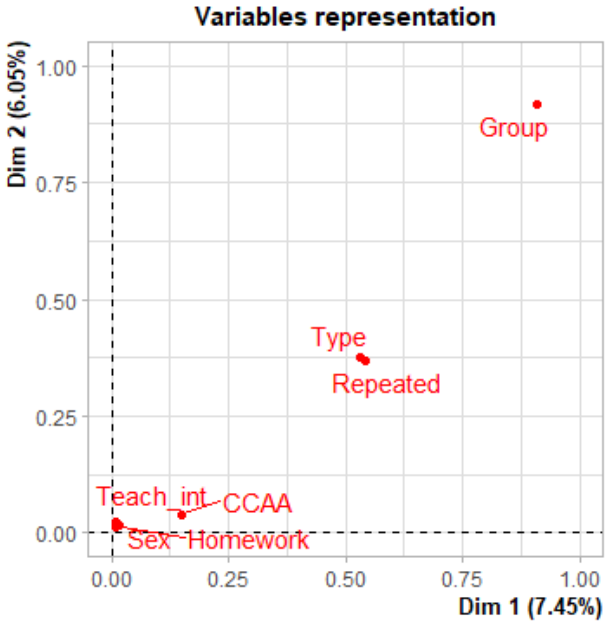




MCA (Multiple Correspondence Analysis)



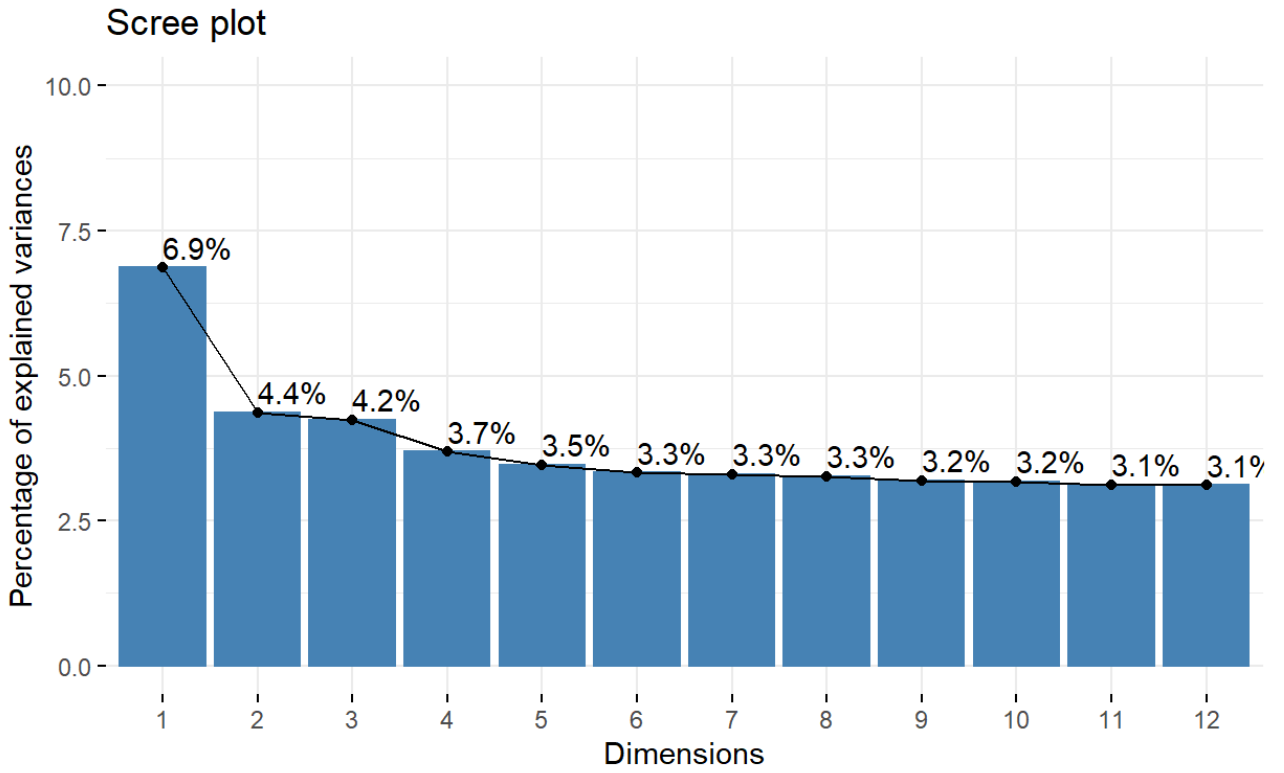
En conclusión, podemos observar que private y público están muy diferenciados, siendo lo más destacable del análisis.
Por otro lado, repeated_yes, once tiene un efecto relevante positivo, al contrario de repeated_no,never cuyo efecto es negativo significativamente.



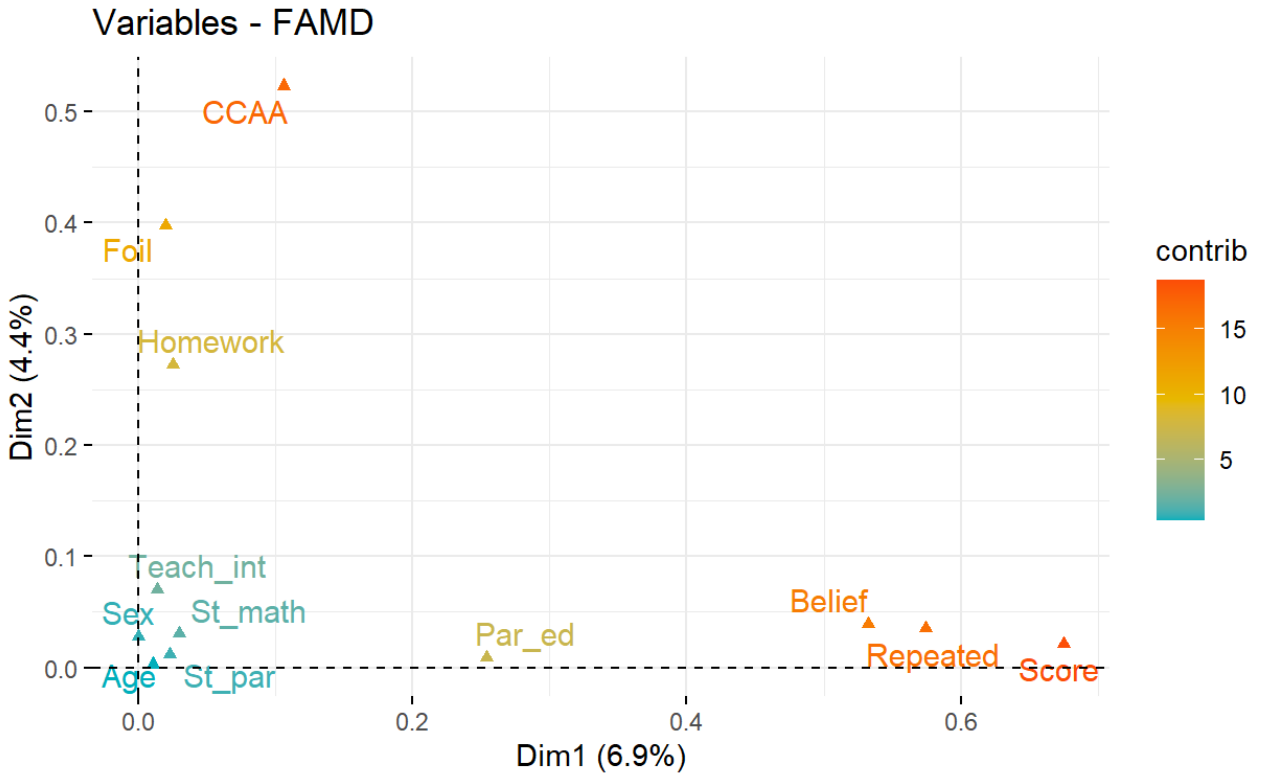
FAMD (Factor Analysis for Mixed Data)

A la hora de realizar los planos factoriales no utilizaremos las variables escuela (School) y estudiante (Student), que son los componentes de la clave primaria de la base de datos, debido a su alto grado de valores únicos.

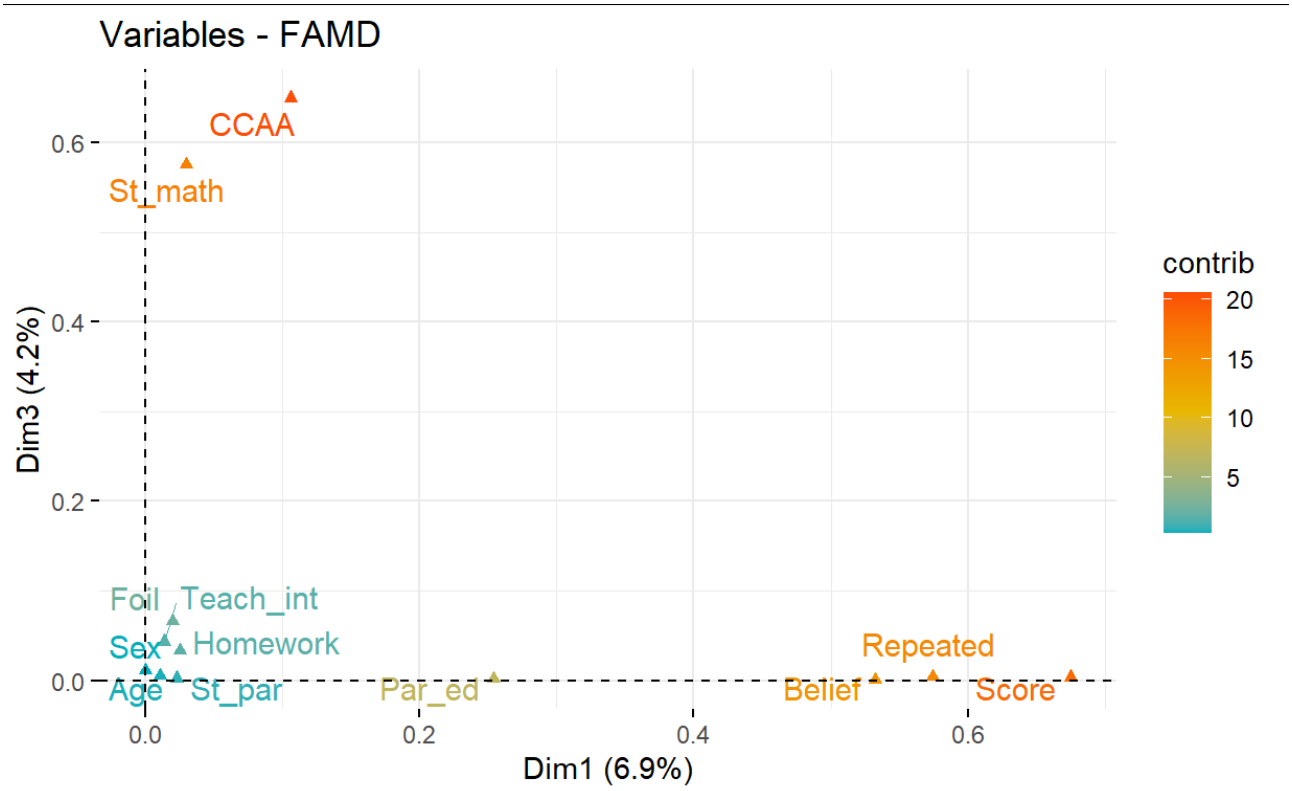
Realizamos un FAMD completo de la base de datos, únicamente siendo suplementaria la variable respuesta (Type).



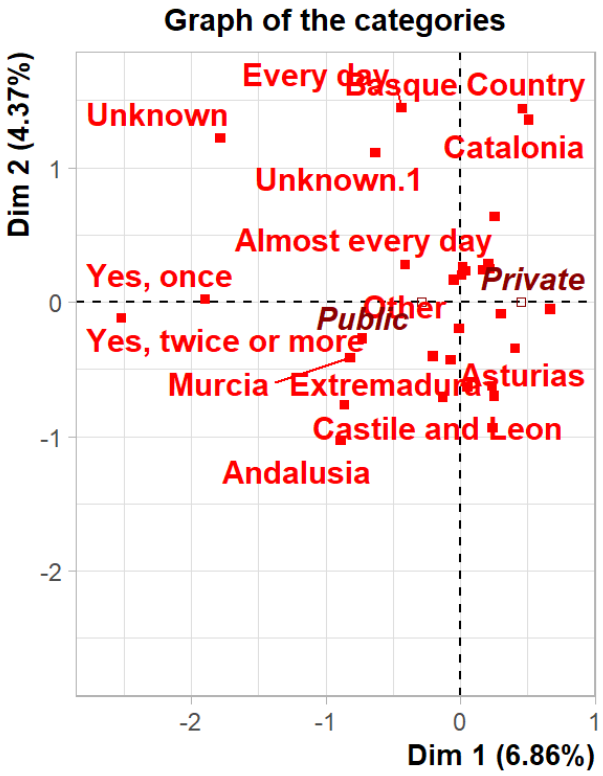
Las primeras 12 dimensiones contienen el 45% de la varianza acumulada. Comprobaremos como se comportan las variables explicativas respecto a los ejes (1,2) y (1,3).



En un principio vemos que en el biplot el primer plano factorial está positivamente relacionado con la nota (Score),la variable categórica de repetición en la ESO (Repeated), la creencia previa de conocimiento en las matemáticas (Belief), y los años máximos de educación alcanzados por alguno de los padres (Par_ed). Mientras que el 2ndo plano factorial está relacionado con la Comunidad Autónoma (CCAA), a la creencia previa de conocimiento en campos de matemáticas no existentes (Foil) y el tiempo dedicado a hacer deberes a la semana(Homework).



El 3er plano factorial esta relacionado con la Comunidad Autónoma (CCAA) y el tiempo dedicado al estudio de las matemáticas (St_math).



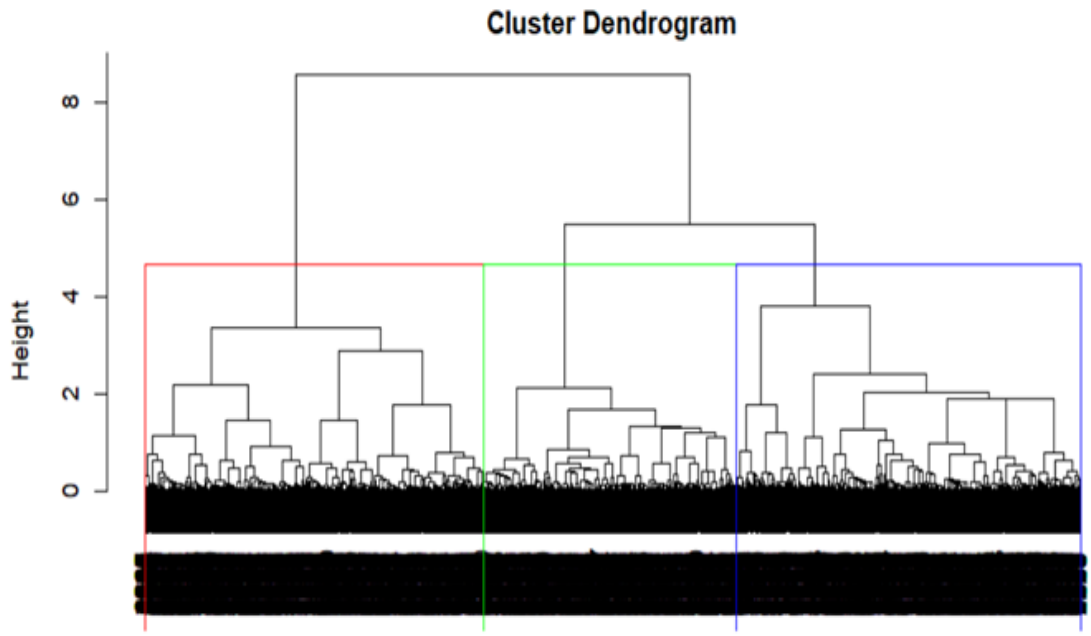
La variable respuesta está relacionada con la primera dimensión, y podemos observar que para la segunda dimensión pueda haber una confusión entre la comunidad autónoma y la lengua del cuestionario (variable no estudiada).

Clustering

El método que hemos utilizado para realizar el clúster es el método de Ward, ya que es un método que funciona bien cuando la base de datos tiene pocas observaciones, en nuestro caso tenemos 8187 observaciones, y una de sus características es que minimiza la varianza dentro de cada clúster.

Para elaborar el análisis de clústeres en datos mixtos (numéricos y categóricos) es necesario establecer una métrica que sea adecuada y en nuestro caso hemos utilizado la distancia de “gower”.

El dendrograma resultante al utilizar la distancia de “gower” es el siguiente:



En el dendrograma vemos claramente que el número de grupos en los que podremos dividir la base de datos será 3 y que todos los clústeres son muy homogéneos entre ellos.

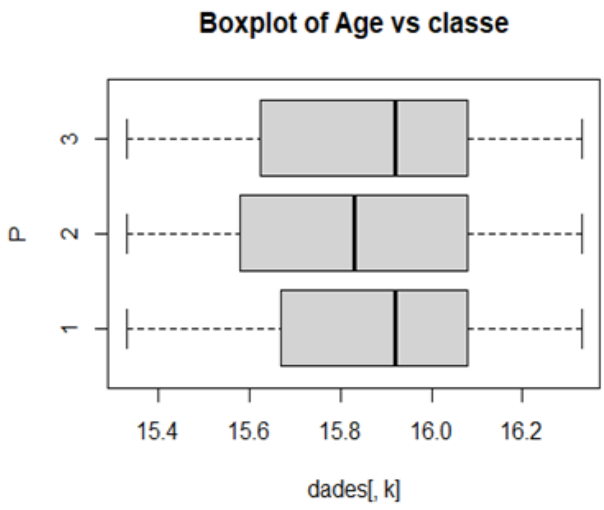
Número de observaciones que tendrá cada clúster:

clúster	1	2	3
nº observaciones	3018	2205	2964

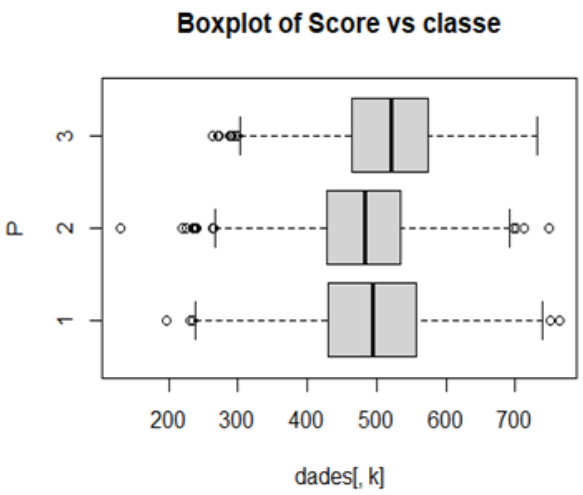
Profiling

Variables Numéricas

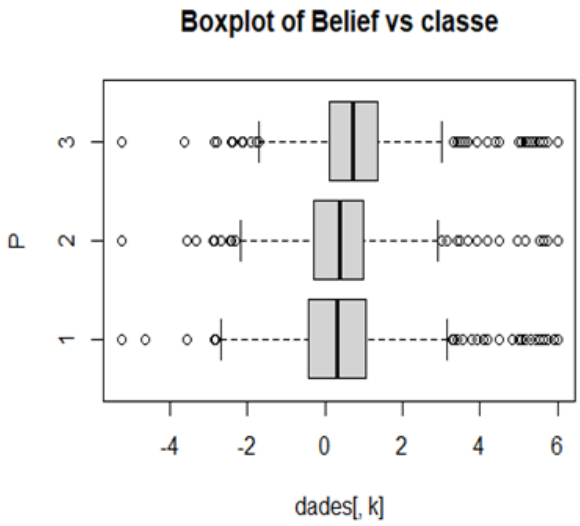
Age:



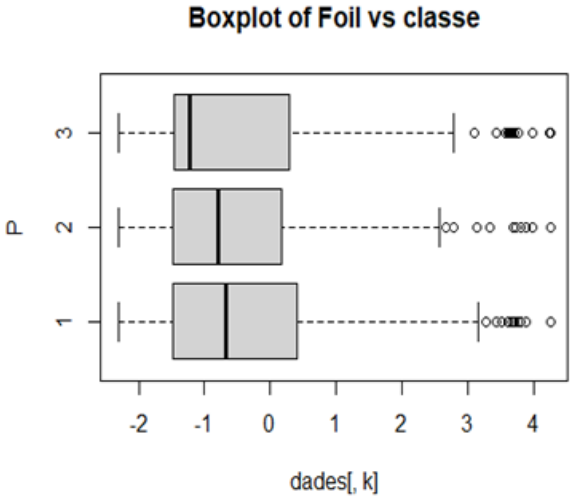
Score:



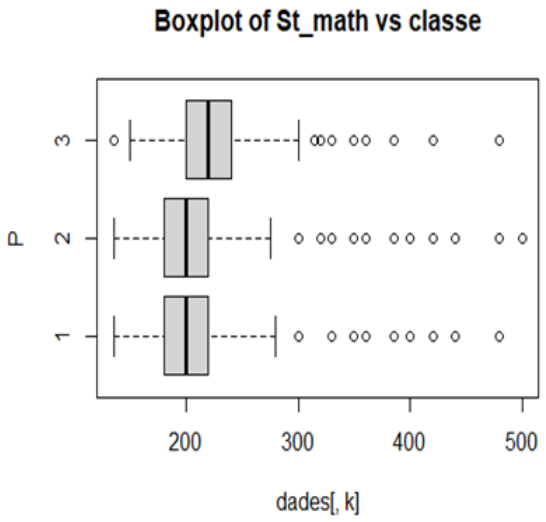
Belief:



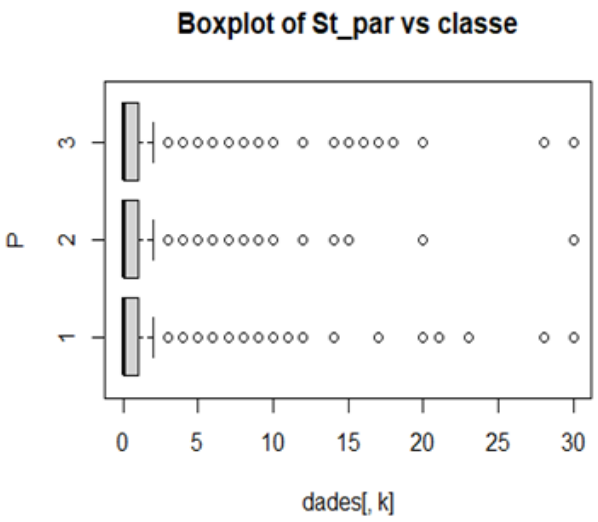
Foil:



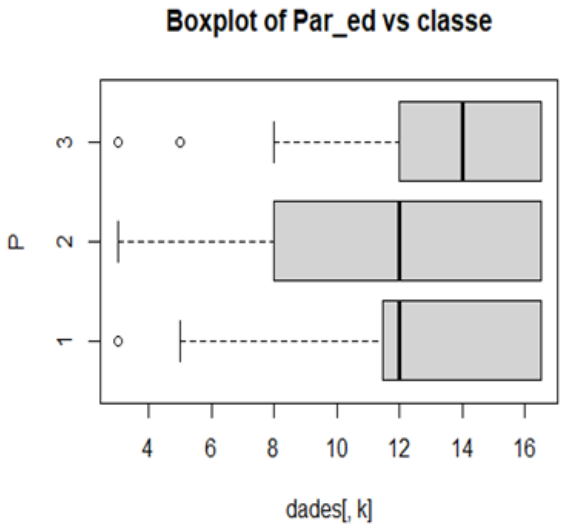
St_math:



St_par:

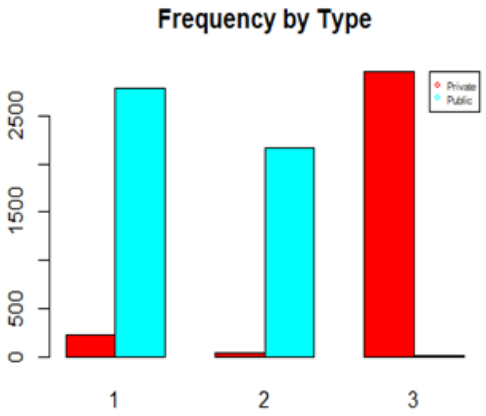


Par_ed:

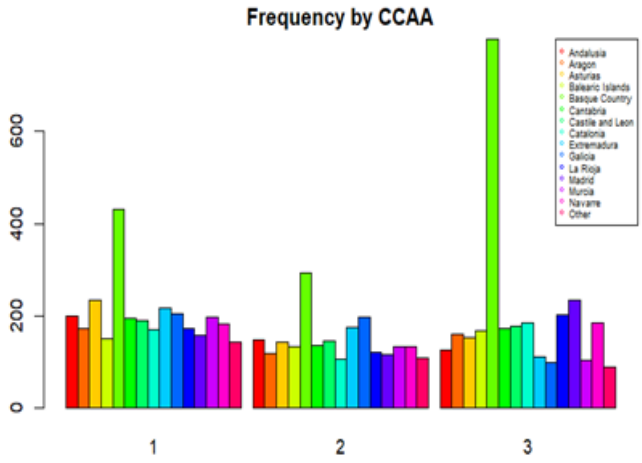


Variables Categóricas

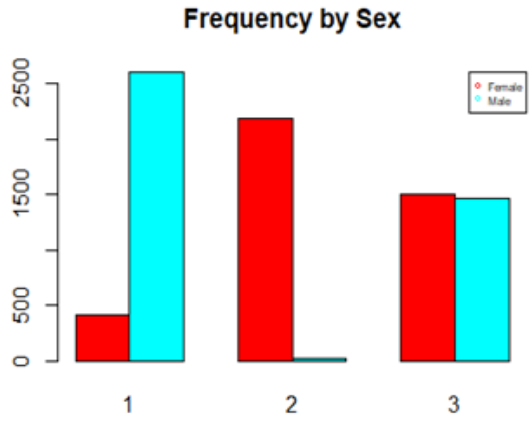
Type:



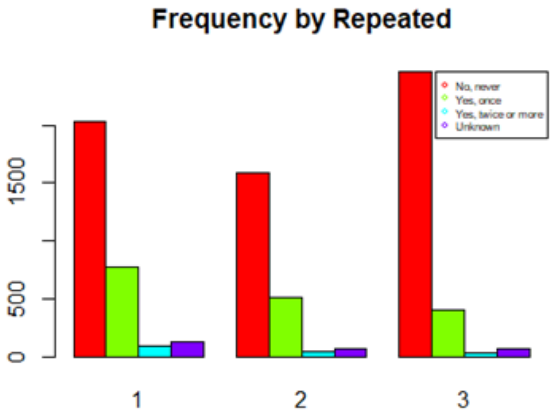
CCAA:



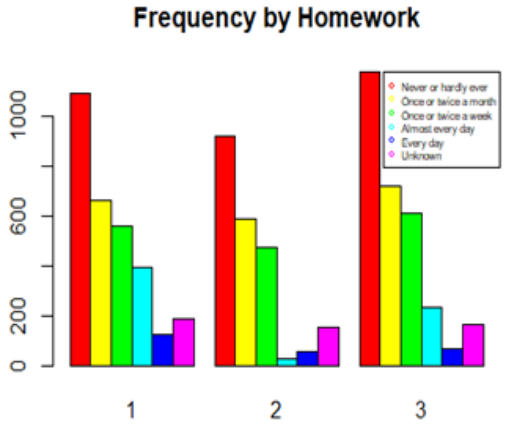
Sex:



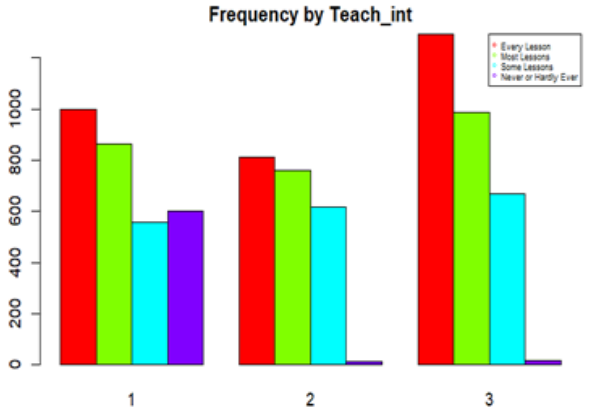
Repeated:



Homework:



Teach_int:



Interpretación general

Se han obtenido resultados para todas las variables estudiadas. En cambio, sólo destacaremos aquellas que hemos considerado más relevantes.

Para las variables numéricas no encontramos ninguna diferencia relevante entre los clústeres.

Por otro lado, en la variable categórica “Type” observamos que los dos primeros clústeres tienen casi el 100% de los valores de las escuelas públicas y el clúster 3 tiene casi todos los estudiantes de escuelas privadas.

Por la variable “Sex” vemos que la mayoría de los estudiantes del clúster 1 son chicos, en el clúster 2 la mayoría son chicas y en el último clúster vemos que los chicos y las chicas están por partes iguales.

De la variable Teach_int podemos ver que todos los estudiantes en los que el docente de matemáticas nunca ha mostrado interés se encuentran en el clúster 1.

Conclusiones

	Clúster 1	**Clúster 2**	**Clúster 3**
Type	Public	Public	Private
Sex	Male	Female	Female/Male
Teach_int	Never or Hardly Ever		

· El clúster 1 representa a todos los alumnos que van a escuelas públicas, son chicos y nunca o casi nunca el profesor ha mostrado interés por ellos. Lo llamamos "**Public_Male_NoInterest**".

· En el clúster 2 encontraremos las chicas que van a escuelas públicas. Lo llamamos "**Public_Female**".

· Por último, en el clúster 3 se encuentran la mayoría de los alumnos de escuelas privadas, independientemente de su sexo. Lo llamamos "**Private**".

Significación

Las variables categóricas son todas significativas.

En las siguientes tablas existe un resumen de las variables numéricas significativas para cada clúster.

Public_Male_NoInterest (clúster 1):

	Age	**Score**	**Belief**	**Foil**	**St_math**	**St_par**	**Par_ed**
Sign.	No	Sí	Sí	Sí	Sí	No	Sí

Public_Female (clúster 2):

	Age	Score	Belief	Foil	St_math	St_par	Par_ed
Sign.	Si	Sí	Sí	Sí	Sí	No	Sí

Private (clúster 3):

	Age	Score	Belief	Foil	St_math	St_par	Par_ed
Sign.	No	Sí	Sí	Sí	Sí	Si	Sí

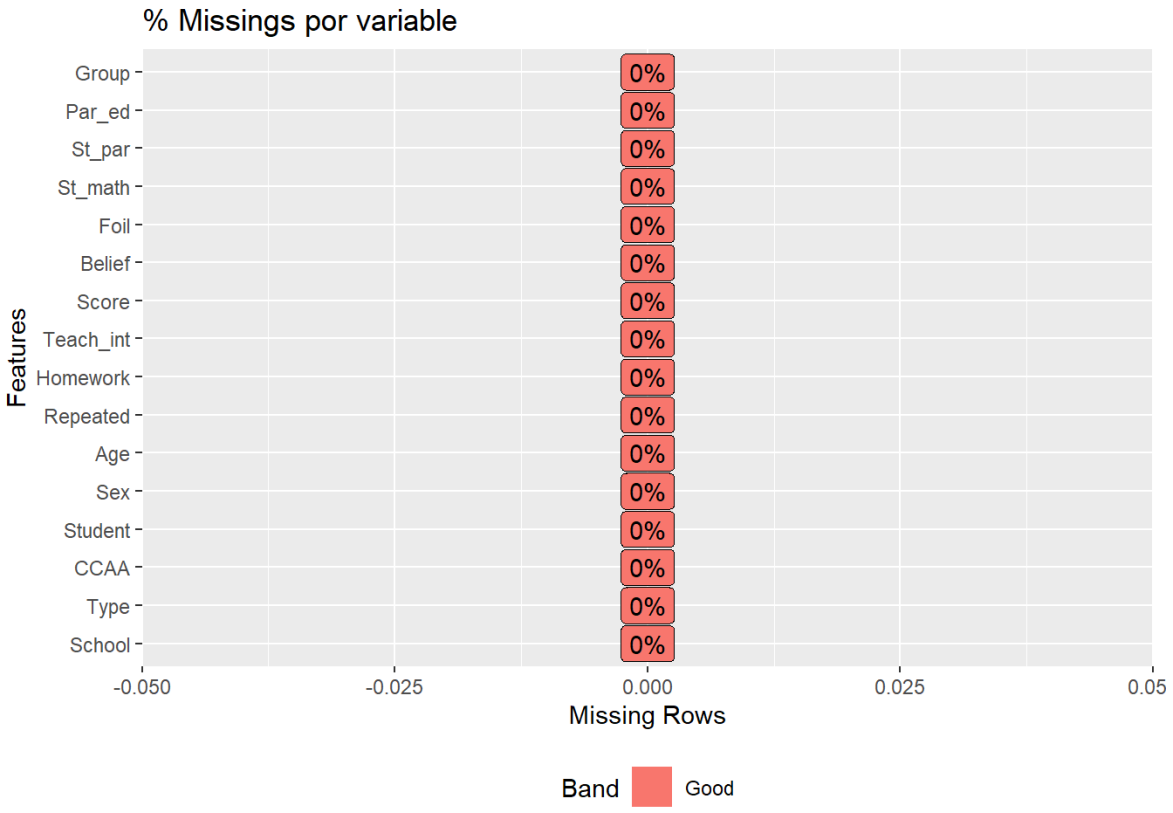
Association Rules

Decision Tree

Un árbol de decisión es un modelo predictivo que usaremos para nuestra base de datos.

Lo primero que debemos hacer es comprobar que no nos quede ninguna variable de las que vayamos a usar con algún valor missing ya que podría provocar problemas y no estaríamos realizando una árbol de decisión adecuado.

En la tabla siguiente observamos que nuestra base de datos ha sido correctamente imputada y que no hay problemas con valores faltantes. La variable Group que vemos aquí es resultado de un clustering previo y no se usará en la construcción del árbol.



- **División y comprobación de equilibrio**

Una vez que hemos comprobado que nuestra base de datos es adecuada, necesitamos dividir nuestra base de datos. La división será necesaria para poder poder proceder con el "entrenamiento" (training) de nuestro árbol y luego poder validarlo con "test". Hay diferentes formas de realizar esta partición, en nuestro caso optamos por elegir ⅔ partes de para training y ⅓ parte para la validación/testeo.

Fijamos una semilla para que los valores no sean diferentes cada vez y ejecutamos el códigos siguiente:

```
# fija una semilla para que no cambie cuando se vuelva a generar el script
set.seed(2108)
#?sample()

#para entrenar
training<-sample(1:nrow(dd2), round(2*nrow(dd2)/3))

#para posterior validacion
test <- dd2[-training, ]
save(test,file ="test.Rdata")
```

Puesto que nuestra variable respuesta es Type, si el colegio es público o privado, tenemos que comprobar que esté bien balanceada en nuestra base de datos. En caso contrario podríamos derivar en errores de sobreajuste (overfitting) y errores de predicción posteriores. En este caso obtenemos que hay un alrededor de un 39% de los datos que corresponden a Private y cerca del 61% que corresponde a Public. Como vemos a continuación:

```
##
##   Private    Public
## 39.36729 60.63271
```

Obtener una base de datos completamente balanceada no suele ser habitual y a pesar de haber métodos para ajustar datos desequilibrados, en nuestro caso acordamos que no dista tanto de ser una base de datos completamente desequilibrada y aceptamos que las 2 categorías están balanceadas.

Para comprobar que al hacer la división de nuestra base de datos no se hayan dividido de forma irregular las 2 categorías, hacemos el cálculo para training y para test. Puesto que obtenemos los mismos porcentajes (40%-60% aprox., ver ANEXO: Decision Tree) continuamos adelante con el proceso de creación del árbol.

- **Construcción del árbol**

Para ello, usamos la función de R rpart, la cúal devuelve un objeto rpart.

```
p1 = rpart(Type ~ ., data=dd2[training,],method="class", parms = list(split="gini"))
save(p1,file = "tree.Rdata")
```

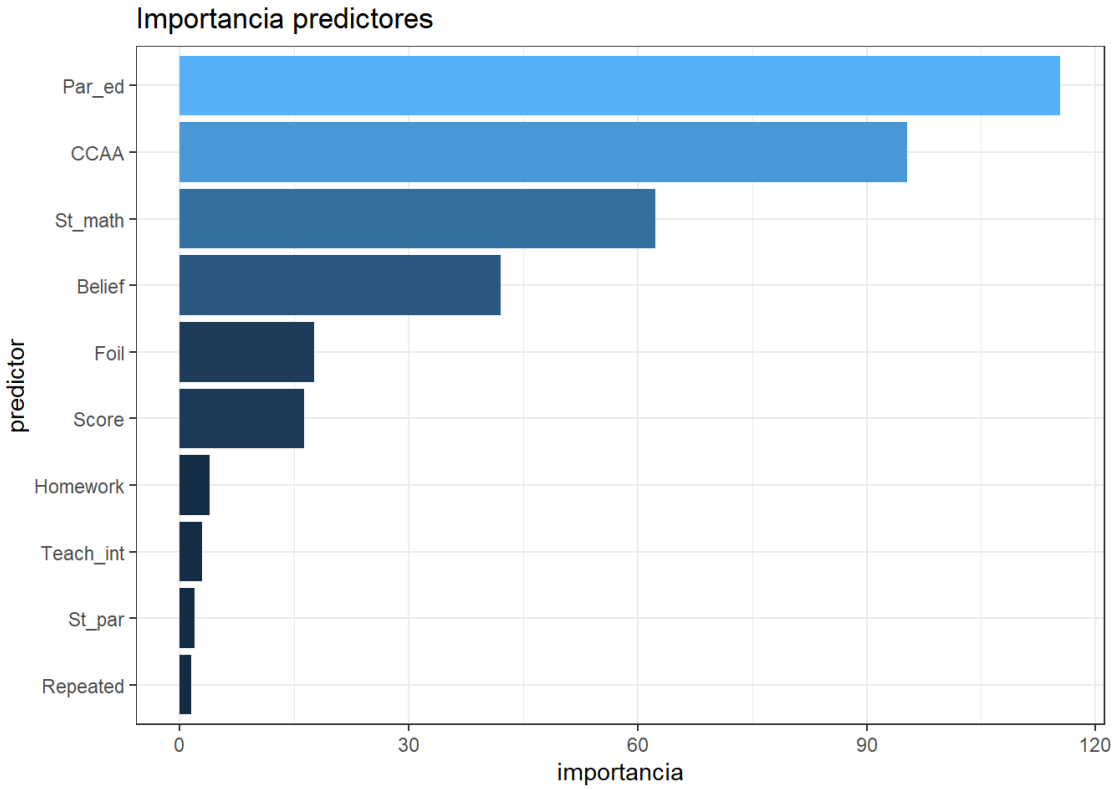
El modelo que usamos para crear el árbol es el siguiente:

Type ~ Repeated + CCAA + Sex + Age + Homework + Teach_int + Score + Belief + Foil + St_math + St_par + Par_ed

Como podemos ver en la función hemos añadido method=”class”, ya que nuestra variable respuesta es categórica. Como método para verificar cuál es la mejor forma de división de las ramas de un árbol, existen varios índices como el “gain information” , “Gini index” o “Entropy”. Nosotros usaremos el índice de Gini.

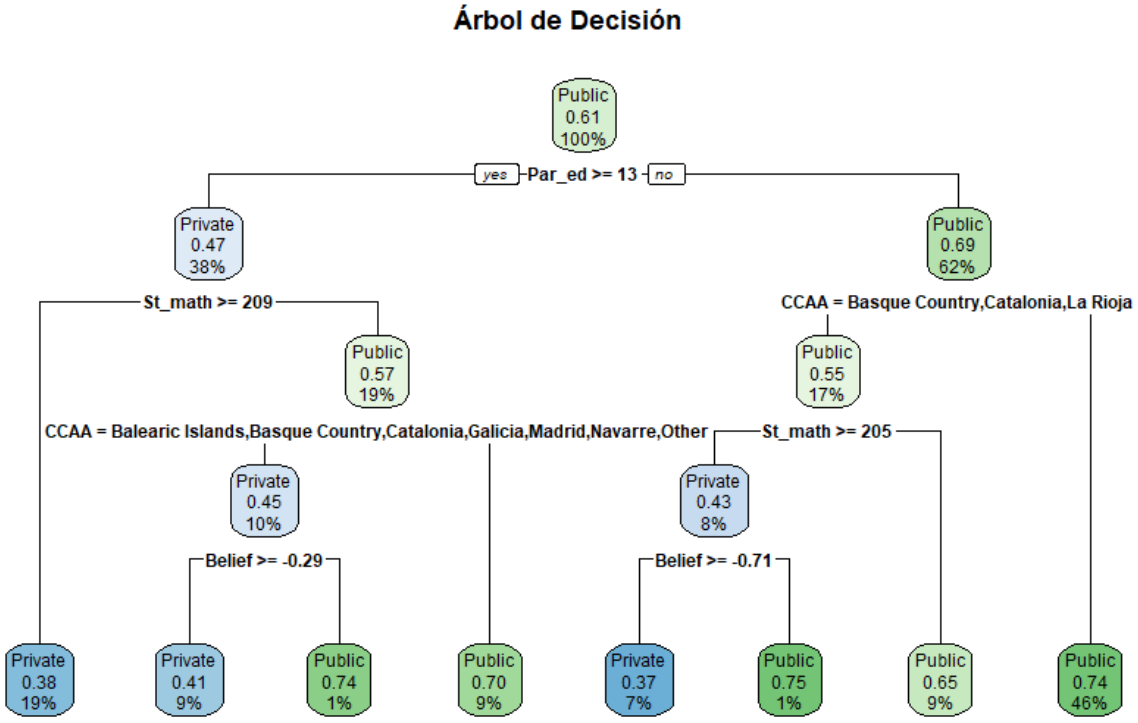
Una vez creado nuestro árbol podemos conocer la importancia que han tenido nuestras variables explicativas en el proceso de creación del árbol (código en ANEXO).

Las cuatro variables más importantes són las comentadas a continuación. La variable más decisiva para crear el árbol es los años de estudio alcanzados por los padres (Par_ed). A continuación tenemos la Comunidad Autónoma a la cual pertenece la escuela (CCAA), seguida del número de horas que dedica el estudiante a estudiar matemáticas (St_math) y finalmente la Media Likert del conocimiento que el estudiante cree tener sobre 10 campos de las matemáticas (Belief). Podemos observar en el gráfico de *Importancia de predictores* todas las variables y su importancia.



- **Plot del árbol de decisión**

En la siguiente imagen podemos ver nuestro árbol de decisión.



De este obtenemos 8 hojas que podemos interpretar de la siguiente forma.

Los caminos que determinarán un resultado de colegio **Público** són:

- Par_ed < 13 & CCAA ≠ País Basco, Catalunya, La Rioja
- Par_ed < 13 & CCAA ≠País Basco, Catalunya, La Rioja & St_math < 205
- Par_ed < 13 & CCAA ≠País Basco, Catalunya, La Rioja & St_math ≥ 205 & Belief < -0.71
- Par_ed ≥ 13 & St_math ≥ 209 & CCAA ≠ Islas Baleares, País Basco, Catalunya, Galicia, Madrid, Navarra y Otros
- Par_ed ≥ 13 & St_math ≥ 209 & CCAA = Islas Baleares, País Basco, Catalunya, Galicia, Madrid, Navarra y Otros & Belief < -0.29

Los caminos que determinarán un resultado de colegio **Privado** són:

- Par_ed < 13 & CCAA ≠País Basco, Catalunya, La Rioja & St_math ≥ 205 & Belief ≥ -0.71
- Par_ed ≥ 13 & St_math ≥ 209 & CCAA = Islas Baleares, País Basco, Catalunya, Galicia, Madrid, Navarra y Otros & Belief ≥ -0.29
- Par_ed ≥ 13 & St_math ≥ 209

- **Precisión**

Para mirar lo bueno que es el modelo resultante, calculamos lo preciso (accuracy) que es nuestro árbol al hacer una predicción tanto para training como para test. Usamos type="class", como hemos hecho anteriormente ya que nuestra variable respuesta es categórica.

```
##           pred
##            Private Public
##    Private    1148    996
##    Public      725   2589
```

```
##          pred_t
##           Private Public
##    Private     567    512
##    Public      393   1257
```

Miramos el ratio de los que ha predicho bien (accuracy): los positivos que ha dado como positivos + los negativos que ha predicho como negativos dividido entre todas las predicciones. Los valores que obtenemos són del 68.57% para los predichos a partir de training y un 66.84% para los valores predichos según test. Ambos valores son bastante similares, por tanto nuestro árbol aunque diste de dar un elevado número de aciertos, supera al menos más del 50% de asertividad.

- Matriz de confusión

Ejecutamos también una matriz de confusión que nos servirá posteriormente para la validación de nuestro árbol.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Private Public
##    Private     567    393
##    Public      512   1257
##
##                Accuracy : 0.6684
##                  95% CI : (0.6504, 0.686)
##     No Information Rate : 0.6046
##     P-Value [Acc > NIR] : 3.303e-12
##
##                   Kappa : 0.2929
##
##  Mcnemar's Test P-Value : 8.765e-05
##
##             Sensitivity : 0.5255
##             Specificity : 0.7618
##          Pos Pred Value : 0.5906
##          Neg Pred Value : 0.7106
##              Prevalence : 0.3954
##          Detection Rate : 0.2078
##    Detection Prevalence : 0.3518
##       Balanced Accuracy : 0.6437
##
##        'Positive' Class : Private
##
```

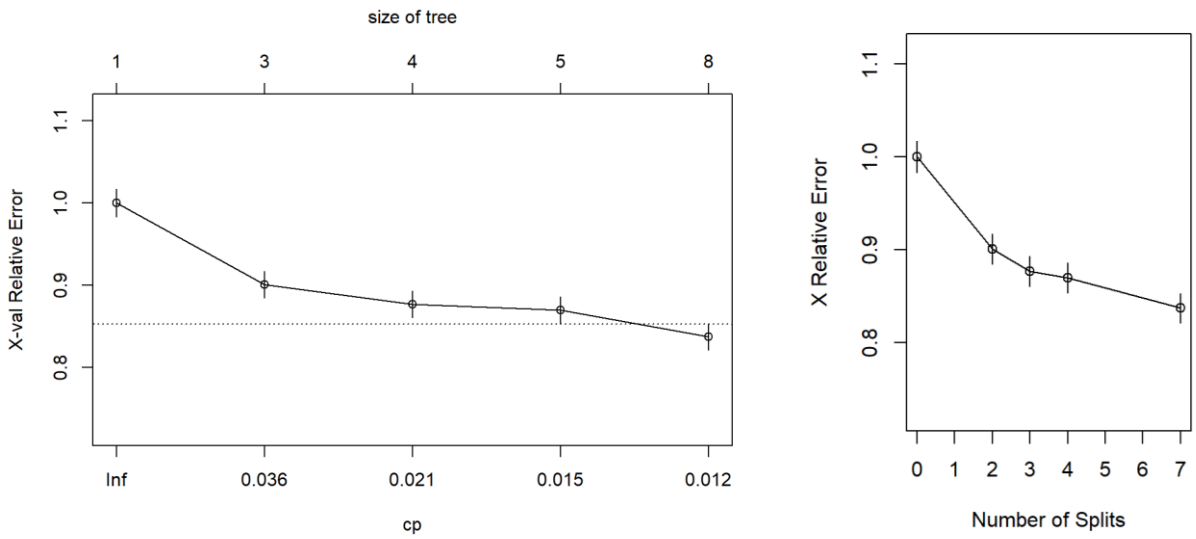
Observamos que tenemos un accuracy de un 67%. Una sensibilidad de un 52% y una especificidad de un 76% aproximadamente. La adecuación de esto la comprobaremos en el apartado de Validations: DECISION TREE. Así pues en este apartado no comentaremos más al respecto.

- **Poda**

Ante la posibilidad de tener en nuestro árbol ramas redundantes y por tanto demasiadas hojas, podemos proceder con el método llamado de “poda”. Para ello usamos la función *prune* de R. Dentro de la cual especificamos que cp (Complexy Parameter) sea el mínimo, es decir, la mínima medida óptima de nuestro árbol de decisión.

```
pfit<- prune(p1, cp = p1$cptable[which.min(p1$cptable[,"xerror"]),"CP"])
```

En nuestro caso obtenemos que nuestro árbol de decisión después de la poda, es el mismo que el anterior a la poda (por tanto el mismo que hemos visto antes). Podemos concluir que el árbol que hemos obtenido usando este método de rpart, ya tenía el tamaño óptimo.



En el gráfico de *size of tree* podemos comprobar que mínimo CP se alcanza con las 8 hojas. Y en el otro gráfico también observamos que el mínimo error relativo se obtiene al hacer 7 cortes.

- Random Forest

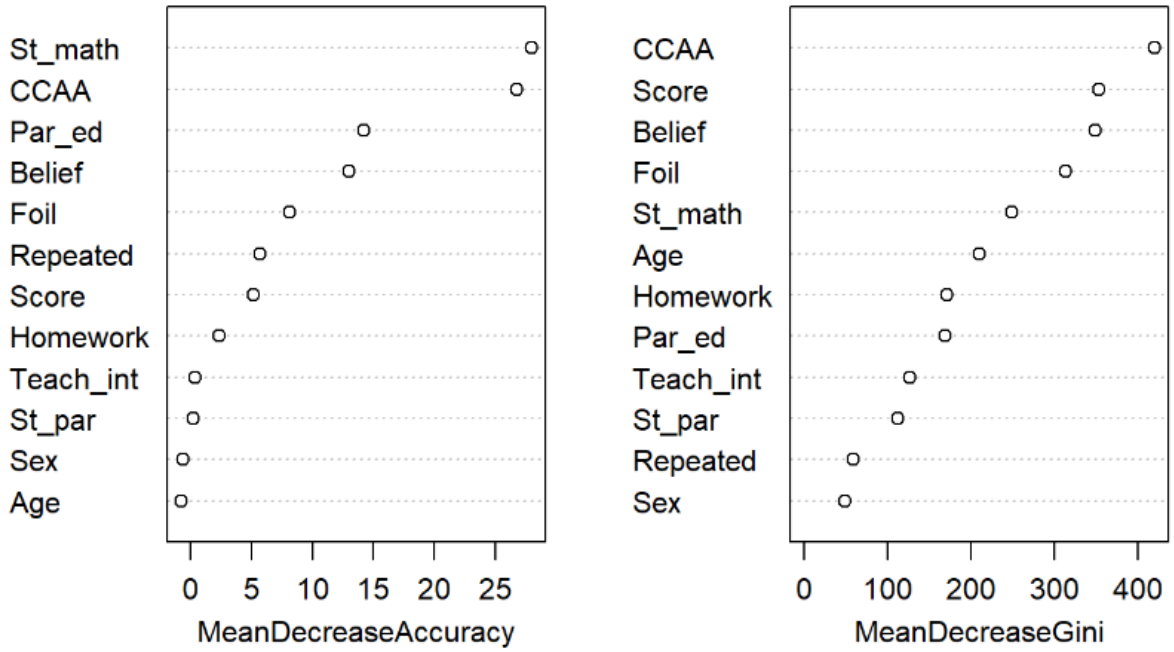
Finalmente ejecutamos también una función que nos permite crear varios árboles a la vez y hacer un cálculo conjunto para tratar de obtener el mejor árbol posible. Hacemos uso de la función randomForest de R.

```
fit4<- randomForest(dd2[training,2:13], y = dd2$Type[training], xtest = dd2[-training,2:13], ytest=dd2
$Type[-training],importance = T, ntree=100, proximity = T, keep.forest = T)
```

Le pedimos que ejecute una cantidad total de 100 árboles y que guarde los resultados. También que calcule la importancia de las variables explicativas (importance=T).

```
## Call:
##  randomForest(x = dd2[training, 2:13], y = dd2$Type[training],      xtest = dd2[-training, 2:13], ytest = dd2
$Type[-training],      ntree = 100, importance = T, proximity = T, keep.forest = T)
##                Type of random forest: classification
##                      Number of trees: 100
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 29.26%
## Confusion matrix:
##         Private Public class.error
## Private    1066   1078   0.5027985
## Public      519   2795   0.1566083
##                Test set error rate: 28.77%
## Confusion matrix:
##         Private Public class.error
## Private     528    551   0.5106580
## Public      234   1416   0.1418182
```

Podemos ver en los siguientes gráficos la importancia de las variables según la media de accuracy. Y respecto Gini.



Al calcular su accuracy tanto usando training como test, obtenemos cerca del 71%. En el apartado de Validations pasaremos a validar nuestros modelos de árbol y randomForest.

LDA

Validations

Para validar y escoger entre nuestros algoritmos de clasificación para la variable respuesta (Type, si la escuela es Pública o Privada) utilizaremos la matriz de confusión y la curva ROC. Todos los métodos han sido entrenados en el mismo training set y validados contra el mismo test set.

DECISION TREE

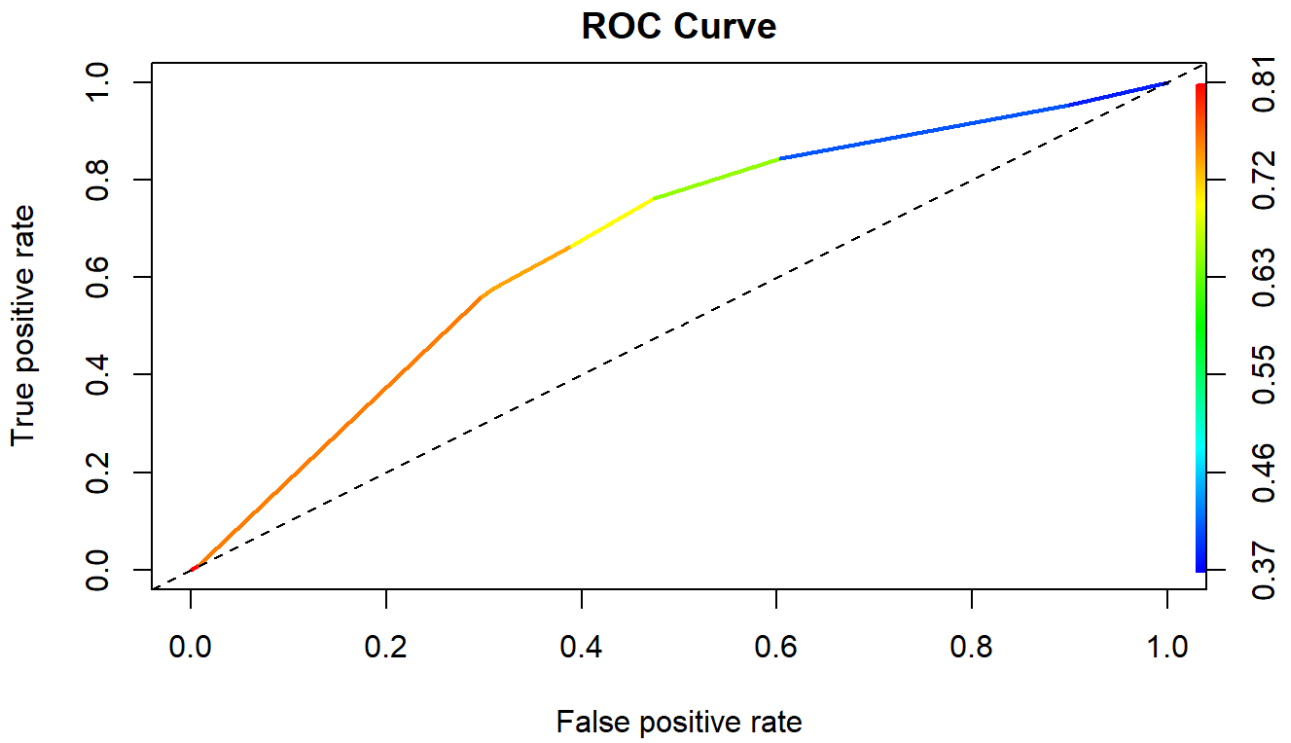
Observamos que las variables que tienen mayor peso en el árbol del apartado previo son Belief (Creencia de conocimientos en matemáticas), St_math (tiempo de estudio de matemáticas) y CCAA.

	Overall <dbl>
Belief	206.205114
CCAA	178.512886
Foil	41.600950
Par_ed	115.143844
Repeated	118.371053
Score	89.662213
St_math	199.444058
St_par	5.335281
Sex	0.000000
Age	0.000000

Nuestro decision tree ha tenido una *accuracy* cercana a los 2/3 (66%), una sensibilidad un poco por encima de la mitad y una especificidad del 76%. Es un resultado bastante pobre para una variable binaria, teniendo en cuenta que tenemos un NIR (No Information Rate) de 0.6046 (Mejor accuracy posible con solo las proporciones de la variable respuesta).

```
Confusion Matrix and Statistics

          Reference
Prediction Private Public
   Private     567    393
   Public      512   1257

               Accuracy : 0.6684
                 95% CI : (0.6504, 0.686)
    No Information Rate : 0.6046
    P-Value [Acc > NIR] : 3.303e-12

                  Kappa : 0.2929

 Mcnemar's Test P-Value : 8.765e-05

            Sensitivity : 0.5255
            Specificity : 0.7618
         Pos Pred Value : 0.5906
         Neg Pred Value : 0.7106
             Prevalence : 0.3954
         Detection Rate : 0.2078
   Detection Prevalence : 0.3518
      Balanced Accuracy : 0.6437

       'Positive' Class : Private
```

Se ejecutó un pruned del árbol, pero no se llegó a cortar ninguna rama. Debido a esto la curva ROC de la predicción del arbol pruned es idéntica a la del árbol previo.

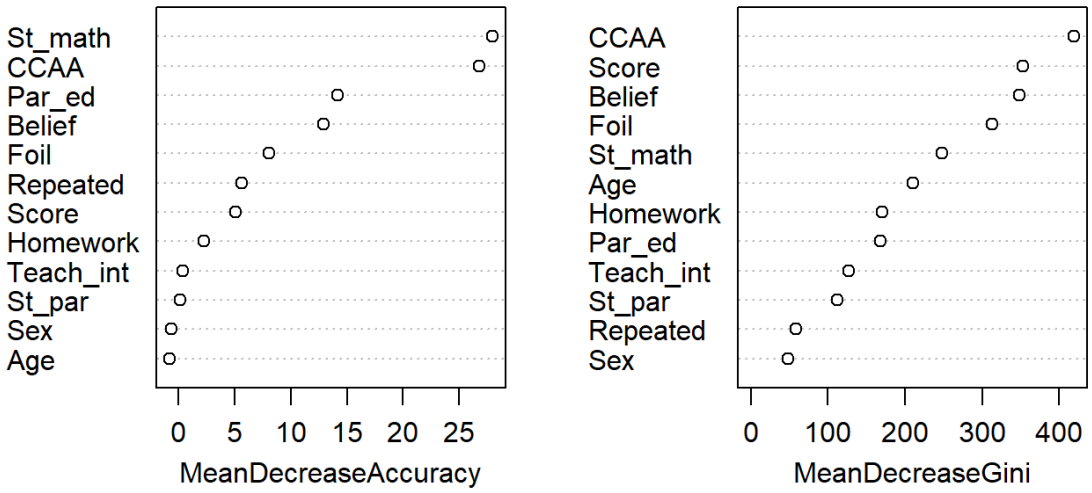


Ahora pasamos a validar el Random Forest generado:

Vemos que St_math,CCAA y Belief siguen siendo de las variables explicativas más importantes también bajo este método.

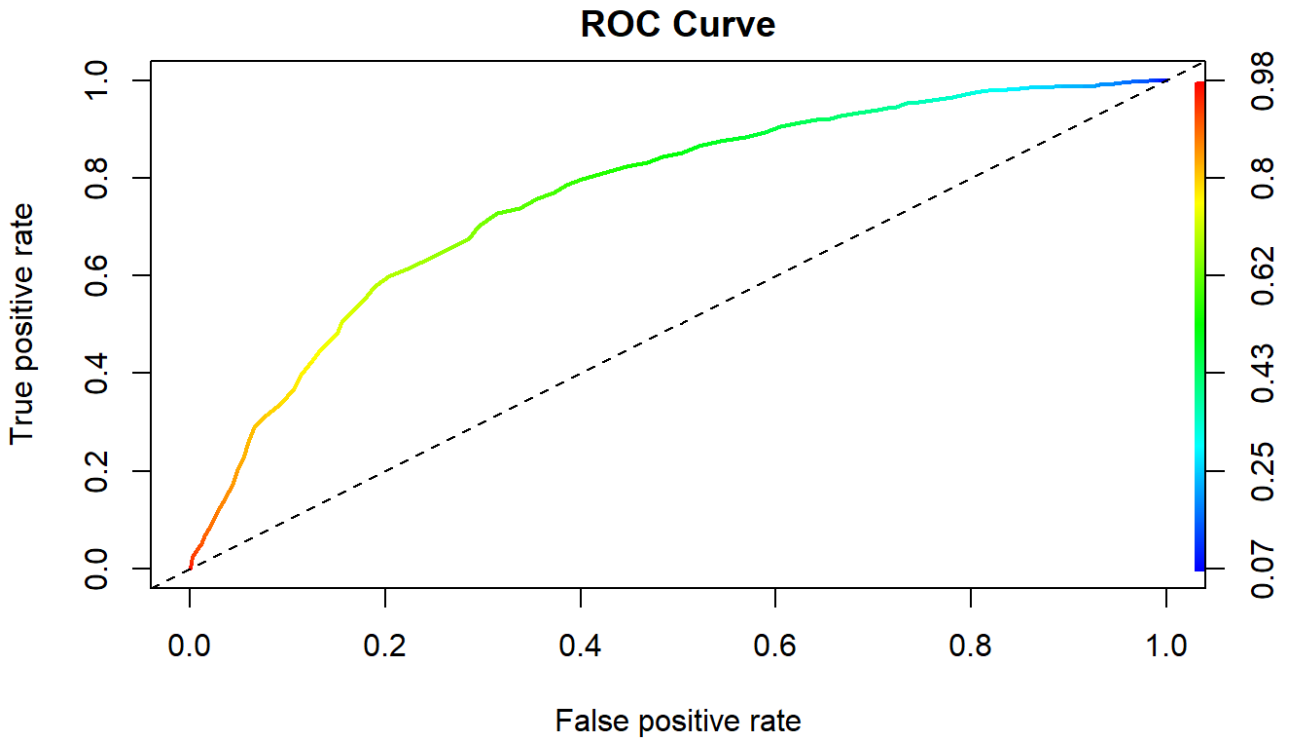
---

fit4




```
Confusion Matrix and Statistics

          Reference
Prediction Private Public
   Private     528    234
   Public      551   1416

               Accuracy : 0.7123
                 95% CI : (0.695, 0.7293)
    No Information Rate : 0.6046
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3661

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.4893
            Specificity : 0.8582
         Pos Pred Value : 0.6929
         Neg Pred Value : 0.7199
             Prevalence : 0.3954
         Detection Rate : 0.1935
   Detection Prevalence : 0.2792
      Balanced Accuracy : 0.6738

       'Positive' Class : Private
```

Gana accuracy y especificidad respecto al previo, pero a cambio de una pérdida de sensibilidad.

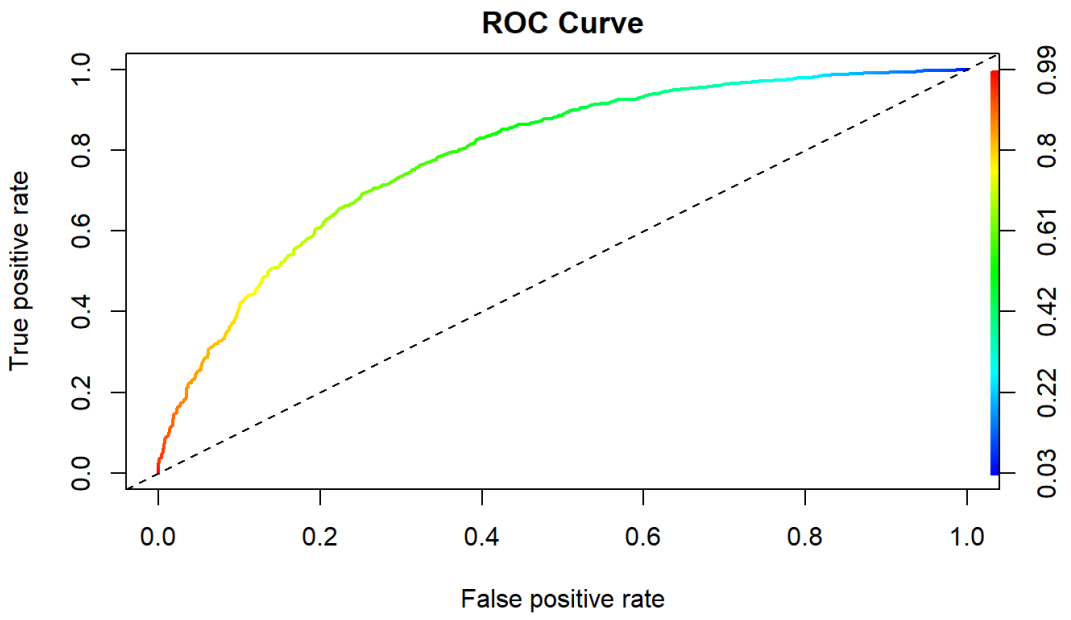


Aunque la curva ROC del RandomForest tiene una mayor área debajo de la curva (AUC) que el Decisions Tree, la pérdida de *accuracy* lleva a que la diferencia en la medida F sea bastante pequeña (RF, 0.5736013 vs DT, 0.556155).

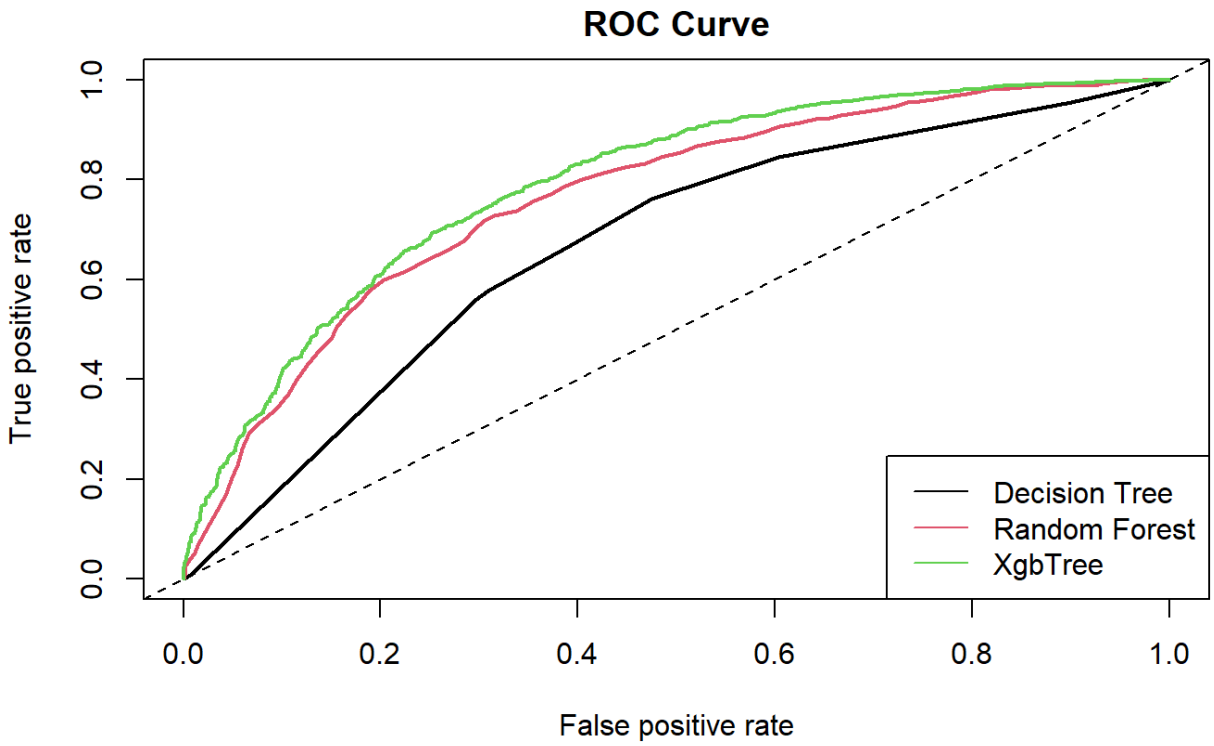
Ahora comprobaremos un método más avanzado con el xgboost (Extreme Gradient Boosting) con crosvalidación; es pasamos de generar un arbol en Decisions Tree ha generar 100 con el Random Forest, pero ahora vamos más allá y generamos 100 rondas de 5 árboles (500), pero iterando al través del método del gradiente (que podemos utilizar ya que a través del paquete caret, xgbtree, recodifica las categóricas en numéricas).

Resampling: Cross-Validated (5 fold), 100 rounds

```
Confusion Matrix and Statistics

          Reference
Prediction Private Public
   Private     593    223
   Public      486   1427

               Accuracy : 0.7402
                 95% CI : (0.7233, 0.7566)
    No Information Rate : 0.6046
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4327

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.5496
            Specificity : 0.8648
         Pos Pred Value : 0.7267
         Neg Pred Value : 0.7459
             Prevalence : 0.3954
         Detection Rate : 0.2173
   Detection Prevalence : 0.2990
      Balanced Accuracy : 0.7072

       'Positive' Class : Private
```



Supera tanto al Random Forest como al Decision Tree en las métricas de la matriz de confusión (F measure: 0.6258575 vs 0.5736013 y 0.556155), y la curva ROC también tiene un mayor AUC (0.791195 vs 0.7643402 y 0.6674828). Pese a esto aún hay un claro margen de mejora para el que sea nuestro modelo predictivo.


ROC Curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Decision Tree
Random Forest
XgbTree

ANEXO

PCA (Principal Component Analysis)

FMA (Functional Mode Analysis)

MCA (Multiple Correspondence Analysis)

Clustering

Association Rules

Decision Tree

Validations and LDA