



Proyecto de Estadística Multivariante: Resultados del pilotaje examen EAES 2020

ESTG1042- ESTADÍSTICA MULTIVARIANTE

PERÍODO: 1T -2021

Guayaquil, Ecuador

Junio del 2021

Elaborado por:

Moisés Alexander Benitez Soriano

moibsoni@espol.edu.ec

Luis Aguilar

lfaguila@espol.edu.ec

Sebastian Francisco Poveda Sandoval

sfpoveda@espol.edu.ec

Dirigido por:

Johnny Pambabay Calero, Ph.D.

PROYECTO DE ESTADÍSTICA MULTIVARIANTE

CONTENIDO

1. Introducción	1
2. Objetivos	1
2.1. Objetivos Generales	1
2.2. Objetivos Específicos	1
3. Fuentes y Metodología	2
4. Análisis Descriptivo de los Datos	2
4.1. Tipo de residencia	2
4.2. Nivel de educación del padre	3
4.3. Aciertos de Matemática	4
4.4. Aciertos de Ciencias Sociales	6
4.5. Sostenimiento del colegio	8
5. Estadística Inferencial	8
5.1. Bondad de ajuste	9
5.2. Pruebas de hipótesis	12
5.3. Análisis de Contingencia	13
6. Estadística Multivariante	15
6.1. Análisis de Correlación Canónico	15
6.2. Análisis de Componentes Principales	22
6.3. Análisis Factorial	30
6.4. Análisis de Conglomerados	45
7. Conclusiones	
8. Bibliografía	

Proyecto PAO 1 - 2021

Grupo 3

22/6/2021

Estadística Multivariante

Introducción

Los datos nos muestran el resultado de una prueba piloto del Examen de Acceso a la Educación Superior (EAES) realizada el 4 de agosto de 2020. En total, se tomaron los resultados de 5737 estudiantes de diversas instituciones educativas del país. Se detalló además varios aspectos generales del estudiante que rindió la prueba, como puede ser el tipo del colegio, la forma de preparación para el examen, nivel de educación de los padres, entre otras variables. Se buscará, mediante el presente documento, determinar cuáles son los factores que determinaron un buen rendimiento en el pilotaje antes mencionado, y que condiciones podrían desempeñar un rol negativo en la preparación del alumno para el EAES.

La prueba “Ser Bachiller”, que pasó a ser llamada “EAES”, es una evaluación que deben rendir obligatoriamente los aspirantes que desean estudiar en una Institución de Educación Superior en el país. Su importancia es tal, que es seguro afirmar que determina el futuro profesional de los recién graduados bachilleres del Ecuador, es por ello que, se realizan pruebas previas al examen o pruebas piloto que permitan determinar si la dificultad, tiempo asignado y otros factores son los justos para que el examen sea rendido en óptimas e iguales condiciones para todos.

Objetivos

Objetivo General:

Determinar los factores como también las variables que influyen en el desempeño académico de los estudiantes de secundaria evaluados en el año 2020 dentro del territorio ecuatoriano y como estos pueden estar relacionados entre sí.

Objetivos Específicos:

- Realizar un análisis descriptivo de los datos basado en las variables de interés sobre el desempeño académico de los alumnos que permita una visualización apropiada y además facilite su comparación e interpretación.
- Aplicar estadística inferencial a los datos con la finalidad de una adecuada extrapolación e inducción, en base a las características de la población estudiada como también de las variables que contribuyen en su educación tales como nivel de educación del padre, provincia, sostenimiento, entre otras.
- Emplear métodos de estadística multivariante para la identificación de las variables que ejercen mayor peso sobre el desempeño académico de la población que es objeto de estudio, así como también posibles factores ajenos a las variables investigadas.

Fuentes y Metodologia

Los datos por usar se dividen dos archivos “.xlsx”, estos son: “*Campos_adicionales_espol*” y “*resultados_pilotaje_04082020_ESPOL*”. Los cuales contienen información sobre estudiantes de diversos colegios del Ecuador, detallando algunos aspectos académicos de su vida, tales como: nivel de educación de sus padres, preparación del estudiante, colegio de origen, etc. Así también se describen los resultados de cada estudiante en una prueba piloto sobre las áreas de matemáticas, literatura, ciencias naturales y ciencias sociales. Por motivos de eficiencia se procedió a unir ambos archivos “.xlsx” en uno solo, conservando la integridad de los datos mediante un “Join” usando como clave primaria el campo de “*Codigo.Estudiante*”.

Las variables en la base en su mayoría son de tipo categóricas como por ejemplo el tipo de vivienda en que reside el estudiante y una de las pocas variables cuantitativas es la cantidad de aciertos totales en prueba piloto.

Como parte del análisis descriptivo se realizará como primer punto un análisis univariante mediante diagrama de barras, histogramas, tablas de frecuencia, medidas de dispersión y de tendencia central, entre otros, a las variables que indican el tipo de residencia de los estudiantes, nivel de educación de los padres y madres, el sostenimiento del colegio origen de los estudiantes y las respuestas correctas por cada campo de conocimiento (*Matemáticas, Literatura, Ciencias Naturales y Ciencias Sociales*); mientras que, en el análisis bivariante o entre mas variables se enfatizará en las variables categóricas y análisis de los resultados del estudiante en la prueba haciendo uso de gráficos de cajas y bigotes así como también de diagramas Raster.

Analisis Descriptivo de los Datos

Tipo de residencia: Para la variable que nos describe el tipo de residencia de los aplicantes, tenemos 9 clasificaciones en la que incluimos: casa o villa, choza o covacha, cuartos en casa de inquilinato, departamento, mediagua, rancho, suite, vivienda colectiva y la opción sin registro. Notemos que el tipo de residencia con mayor frecuencia es el de casa o vivienda (54,94%), seguido de sin registro, lo cual nos da a entender que muchos de los estudiantes omitieron ésta pregunta u olvidaron responderla. El porcentaje de aplicantes que viven en un departamento es de 8% aproximadamente, el tercer valor más elevado.

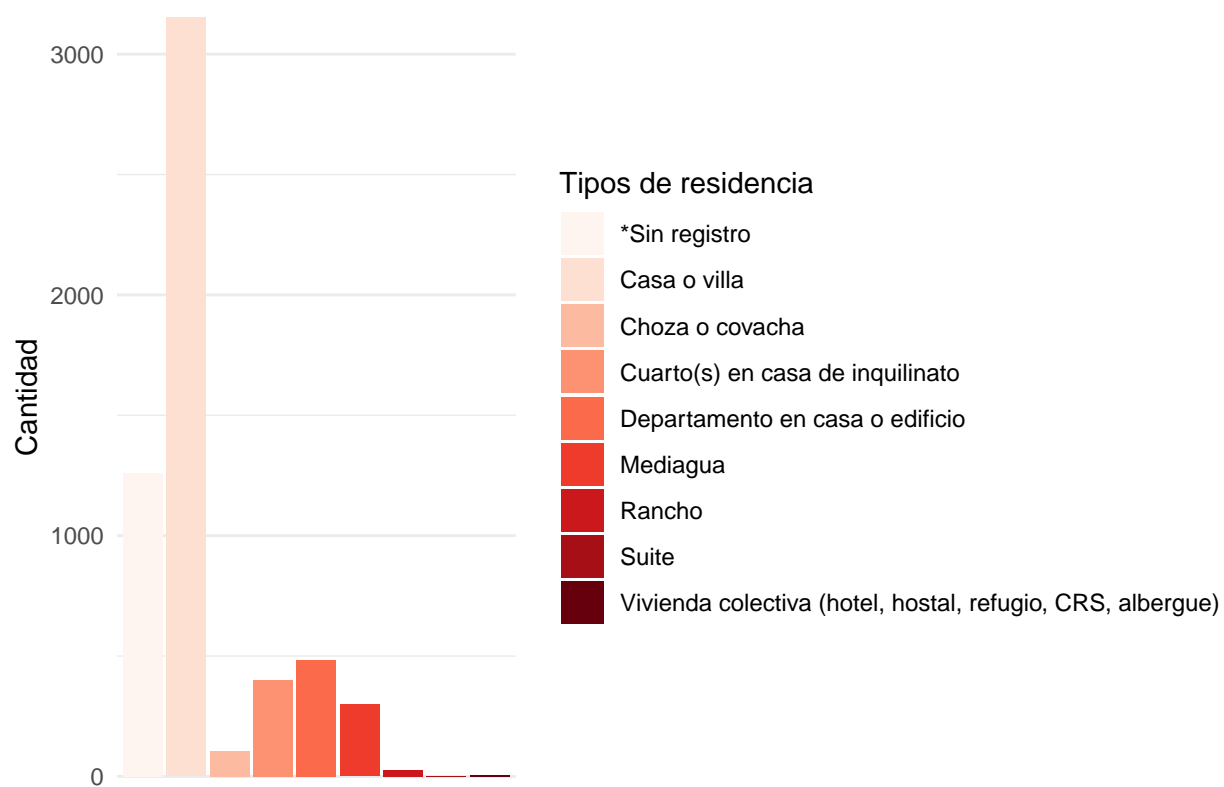


Table 1: Tabla de frecuencias para el tipo de residencia

	Tipo de residencia	Frecuencia	Frecuencia acumulada	Frecuencia relativa	Frecuencia relativa acumulada
1	*Sin registro	1259	1259	0,22	0,22
2	Casa o villa	3152	4411	0,55	0,77
3	Choza o covacha	107	4518	0,02	0,79
4	Cuarto(s) en casa de inquilinato	400	4918	0,07	0,86
5	Departamento en casa o edificio	484	5402	0,08	0,94
6	Mediagua	299	5701	0,05	0,99
7	Rancho	27	5728	0,00	1,00
8	Suite	2	5730	0,00	1,00
9	Vivienda colectiva (hotel, hostel, refugio, CRS, albergue)	7	5737	0,00	1,00

Nivel de educación del padre: La variable “Nivel de educación del padre”, tiene 12 categorías que son descritas perfectamente en la figura. Aunque el porcentaje de sin registro sigue siendo elevado, podemos ver una mejor distribución de los datos en varias de las categorías. Podemos notar que un buen porcentaje de los padres (27.62%) terminó el bachillerato, sin embargo, solo el 11.05% del total cuenta con un título de tercer nivel o superior. Sumando las respuestas no sé y sin registro se tiene un 30.14% de los padres que se desconoce su nivel de educación.

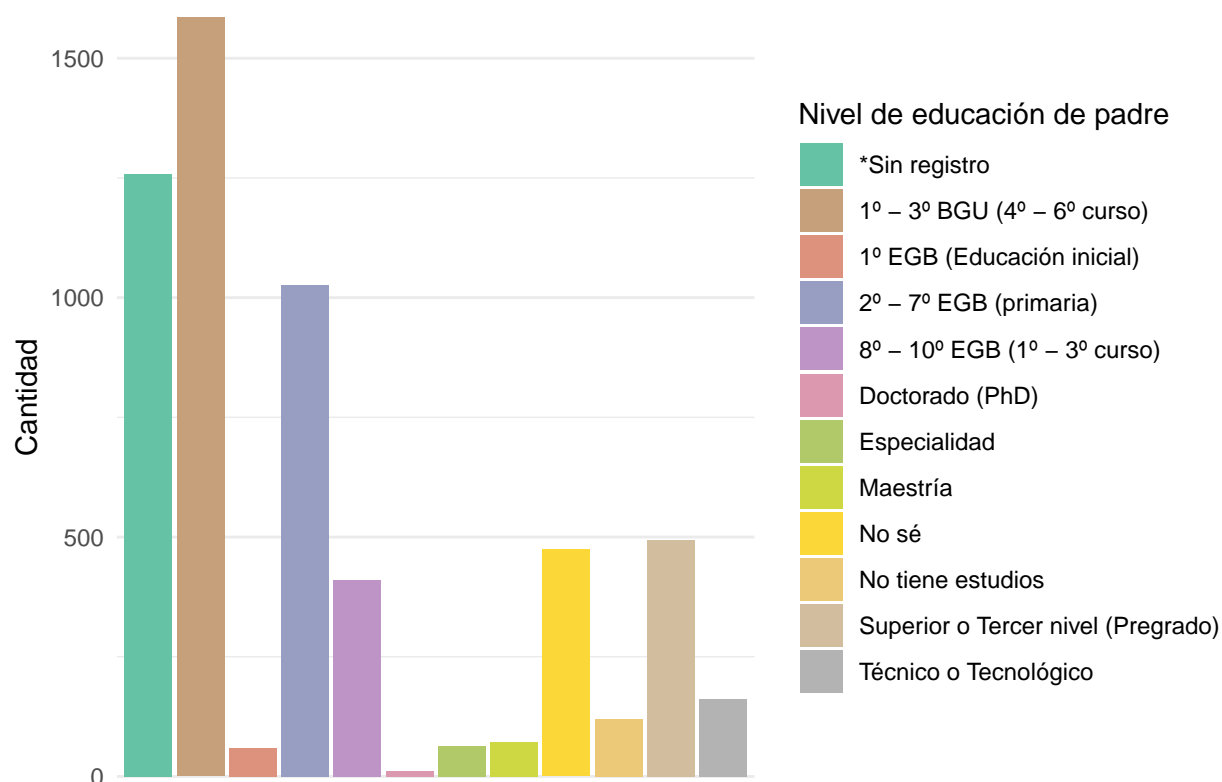


Table 2: Tabla de frecuencias para el nivel de educación del padre

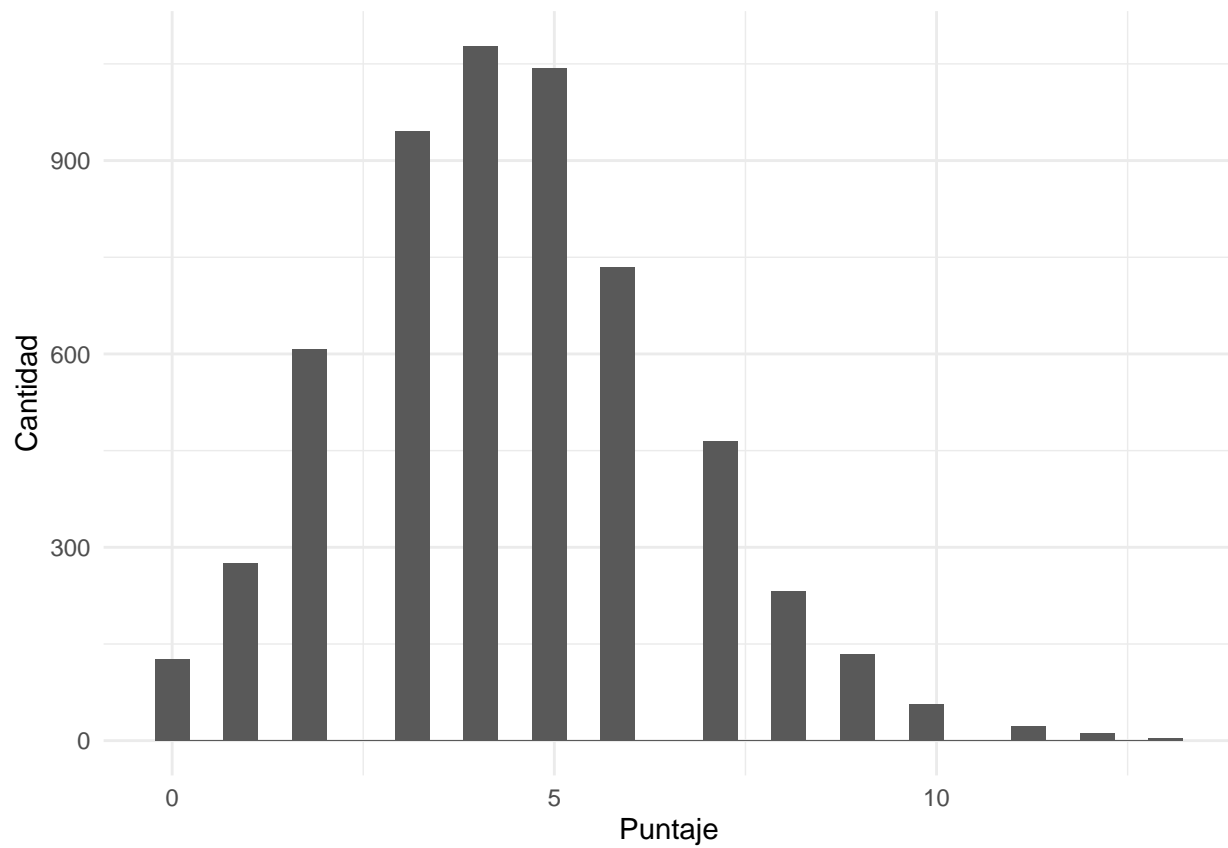
	Nivel de educación	Frecuencia	Frecuencia acumulada	Frecuencia relativa	Frecuencia relativa acumulada
1	*Sin registro	1259	1259	0,22	0,22
2	1º - 3º BGU (4º - 6º curso)	1585	2844	0,28	0,50
3	1º EGB (Educación inicial)	60	2904	0,01	0,51
4	2º - 7º EGB (primaria)	1026	3930	0,18	0,69
5	8º - 10º EGB (1º - 3º curso)	411	4341	0,07	0,76
6	Doctorado (PhD)	12	4353	0,00	0,76
7	Especialidad	64	4417	0,01	0,77
8	Maestría	72	4489	0,01	0,78
9	No sé	474	4963	0,08	0,87
10	No tiene estudios	120	5083	0,02	0,89
11	Superior o Tercer nivel (Pregrado)	493	5576	0,09	0,97
12	Técnico o Tecnológico	161	5737	0,03	1,00

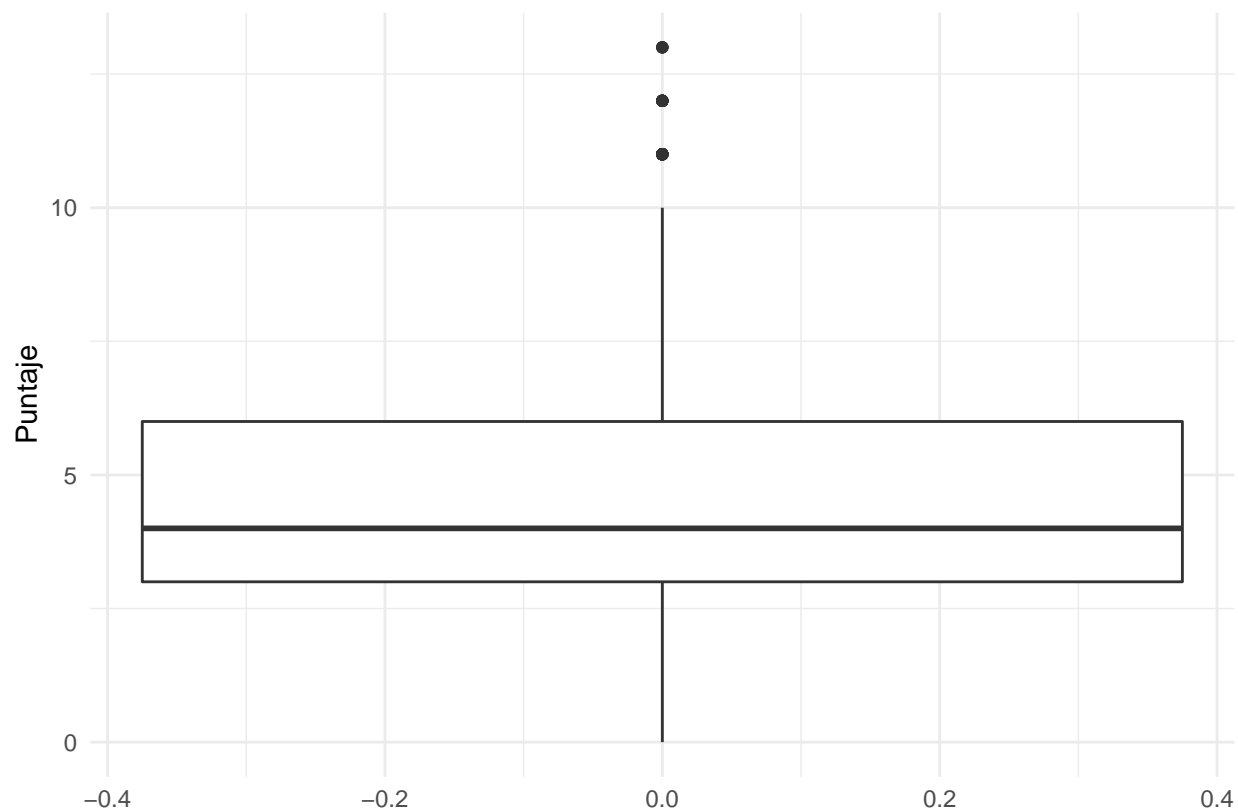
Aciertos de Matemática: Como se puede observar la mayoría de estudiantes obtuvieron una calificación entre 3 y 6 puntos en esta asignatura, teniendo una media de 4.46 pero dado que es una variable discreta la mediana es de 4 y tiene pequeños valores atípicos en 11, 12 y 13 puntos.

Table 3: Tabla de frecuencias para aciertos de Matematicas

	Puntaje	Frecuencia	Frecuencia acumulada	Frecuencia relativa	Frecuencia relativa acumulada
1	0	126	126	0,02	0,02
2	1	276	402	0,05	0,07
3	2	608	1010	0,11	0,18
4	3	945	1955	0,16	0,34
5	4	1078	3033	0,19	0,53
6	5	1043	4076	0,18	0,71
7	6	735	4811	0,13	0,84
8	7	464	5275	0,08	0,92
9	8	232	5507	0,04	0,96
10	9	134	5641	0,02	0,98
11	10	57	5698	0,01	0,99
12	11	23	5721	0,00	1,00
13	12	12	5733	0,00	1,00
14	13	4	5737	0,00	1,00

'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.





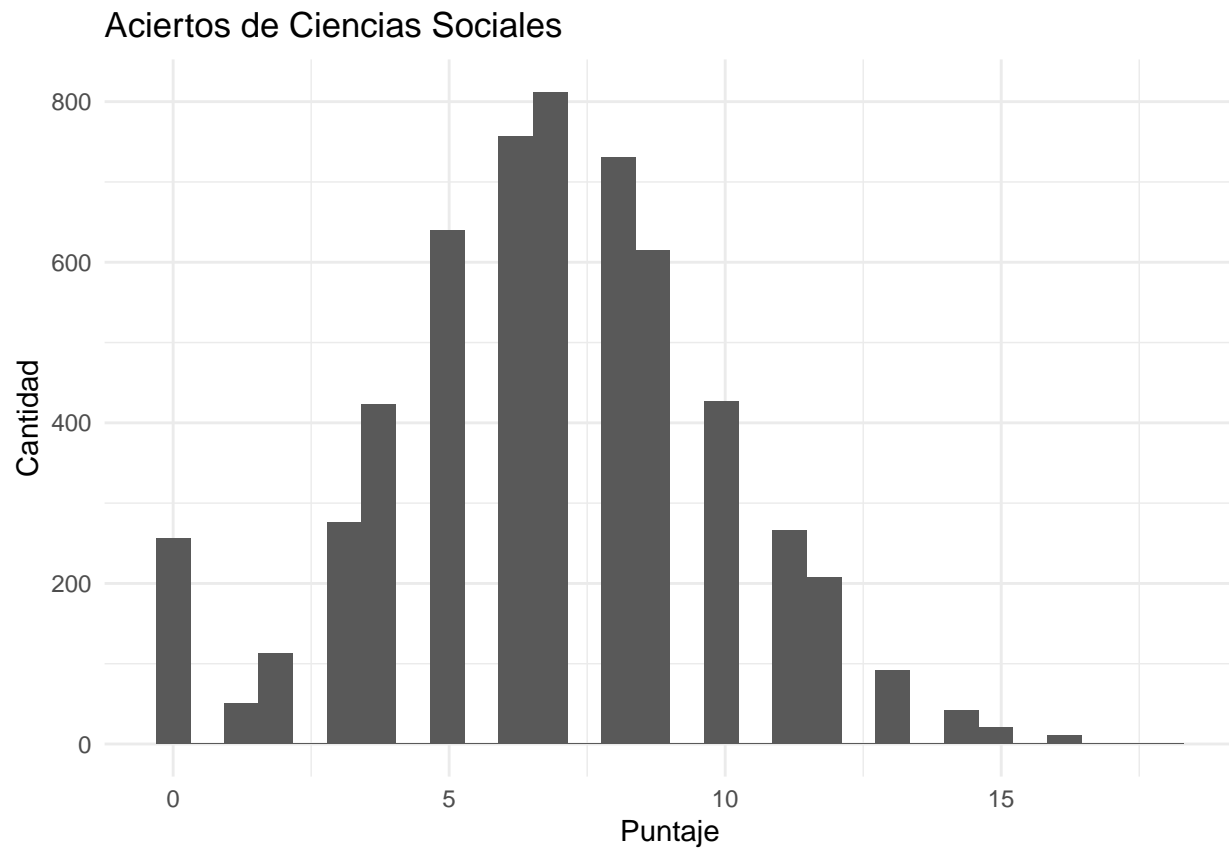
Aciertos de Ciencias Sociales: Se aprecia que gran parte de los estudiantes obtuvieron una calificación entre 5 y 9 puntos en esta asignatura, teniendo una media de 6.89 pero dado que es una variable discreta la mediana es de 7 y tiene bastantes valores atípicos que van desde 15 hasta 18 puntos.

Table 4: Tabla de frecuencias para aciertos de Ciencias Sociales

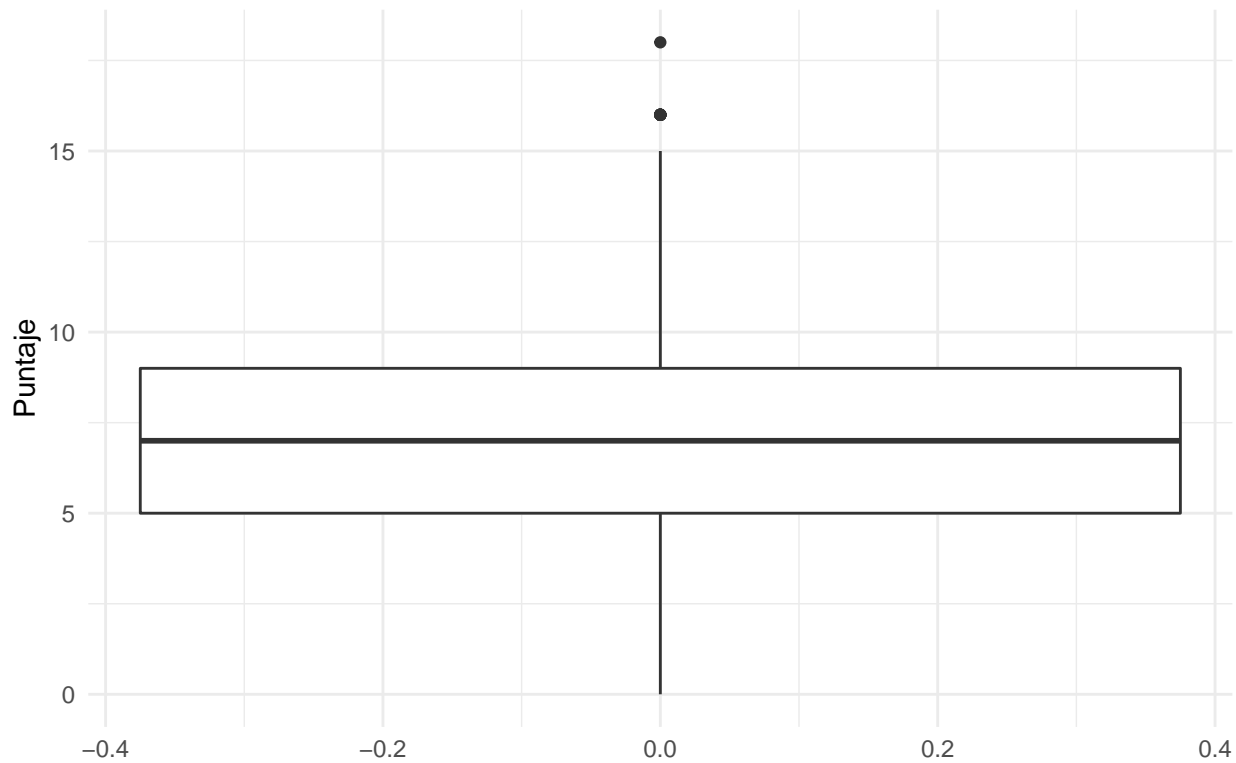
	Puntaje	Frecuencia	Frecuencia acumulada	Frecuencia relativa	Frecuencia relativa acumulada
1	0	256	256	0,04	0,04
2	1	51	307	0,01	0,05
3	2	113	420	0,02	0,07
4	3	276	696	0,05	0,12
5	4	423	1119	0,07	0,20
6	5	640	1759	0,11	0,31
7	6	756	2515	0,13	0,44
8	7	812	3327	0,14	0,58
9	8	731	4058	0,13	0,71
10	9	615	4673	0,11	0,81
11	10	427	5100	0,07	0,89
12	11	266	5366	0,05	0,94
13	12	207	5573	0,04	0,97
14	13	91	5664	0,02	0,99
15	14	42	5706	0,01	0,99
16	15	20	5726	0,00	1,00
17	16	10	5736	0,00	1,00

	Puntaje	Frecuencia	Frecuencia acumulada	Frecuencia relativa	Frecuencia relativa acumulada
18	18	1	5737	0,00	1,00

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



Aciertos de Ciencias Sociales

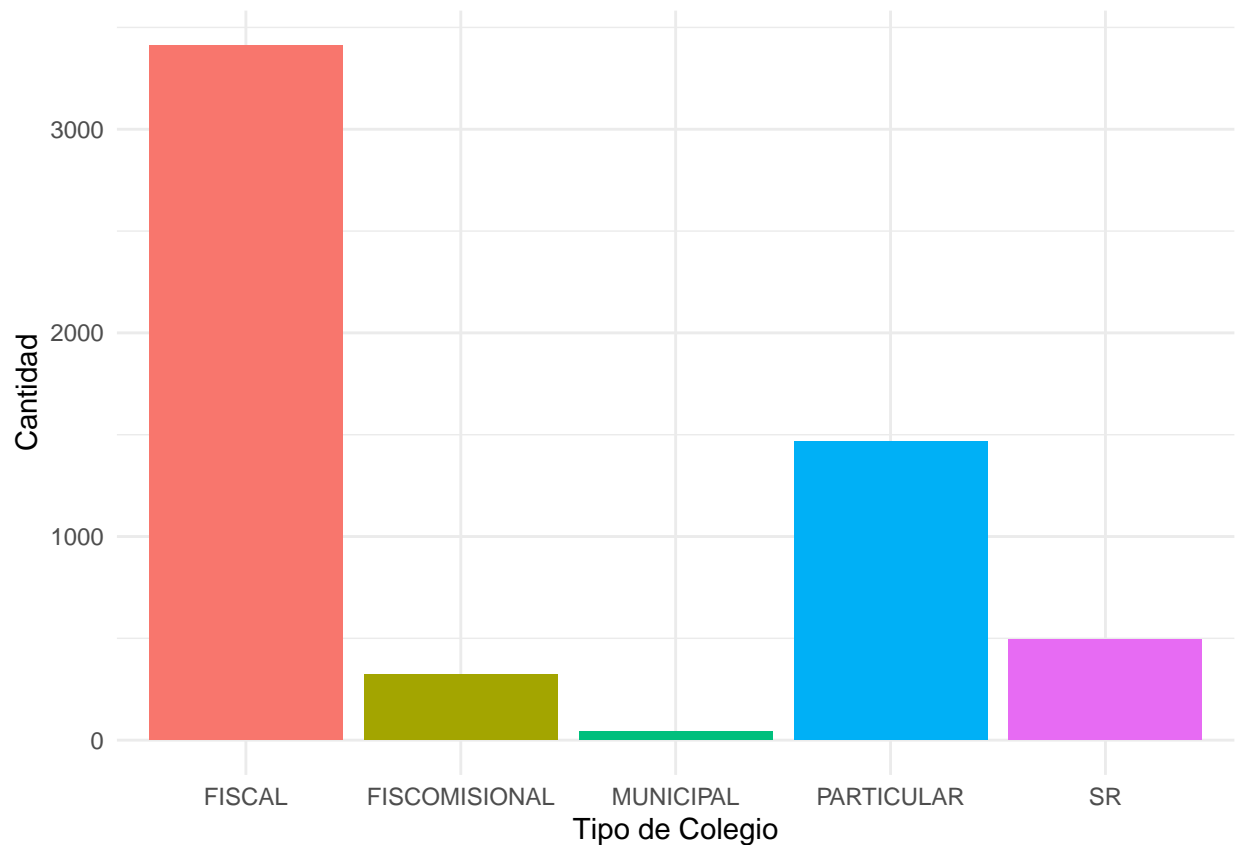


Por una simple inspeccion a breves rasgos de manera visual podemos apreciar que la distribucion de los puntajes por cada materia se asemejan a una normal y aún mas si el tamaño de la muestra fuese mucho mayor.

Sostenimiento del colegio origen de los estudiantes: Como resultados tenemos que el 59% de los estudiantes provienen de un colegio de tipo **FISCAL** en oposicion a los colegios de tipo **PARTICULAR** que es de un 26%, esto se puede deber a algun factor de tipo socioeconomico que desconocemos; mientras que, cerca del 1% son de tipo **MUNICIPAL**.

Table 5: Tabla de frecuencias para el sostenimiento del colegio origen

	Puntaje	Frecuencia	Frecuencia acumulada	Frecuencia relativa	Frecuencia relativa acumulada
1	FISCAL	3412	3412	0,59	0,59
2	FISCOMISIONAL	321	3733	0,06	0,65
3	MUNICIPAL	42	3775	0,01	0,66
4	PARTICULAR	1467	5242	0,26	0,91
5	SR	495	5737	0,09	1,00



Estadística Inferencial

Bondad de ajuste

Se quiere analizar que distribución siguen los datos de la variable *Aciertos Totales*, primeramente planteemos el contraste de hipótesis:

$$H_0 : X \sim N(u, \sigma^2) \quad H_1 : \text{Los datos no siguen una distribución Normal}$$

Estimación de μ , σ^2 y demás:

```
library(psych)
#Media
media= mean(datos$`Aciertos Totales`)
media
```

```
## [1] 29.40753
```

```
#Varianza
varianza = var(datos$`Aciertos Totales`)
varianza
```

```
## [1] 83.45906
```

```
#Desviación
desv = sqrt(varianza)
desv
```

```
## [1] 9.135593
```

```
#Mediana y moda
median(datos$`Aciertos Totales`)
```

```
## [1] 29
```

```
mfv(datos$`Aciertos Totales`)
```

```
## [1] 28
```

```
#Summary
summary(datos$`Aciertos Totales`)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.00   24.00   29.00   29.41   35.00   75.00
```

```
#Sesgo y curtosis directamente estandarizado
skew(datos$`Aciertos Totales`)/sqrt(6/5737)
```

```
## [1] -5.53591
```

```
kurtosi(datos$`Aciertos Totales`)/sqrt(6/5737)
```

```
## [1] 37.64458
```

Si plotamos los datos para revisar la función de densidad de los datos:

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
##
```

```
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
```

```
##
```

```
##      recode
```

```
## The following object is masked from 'package:psych':
```

```
##
```

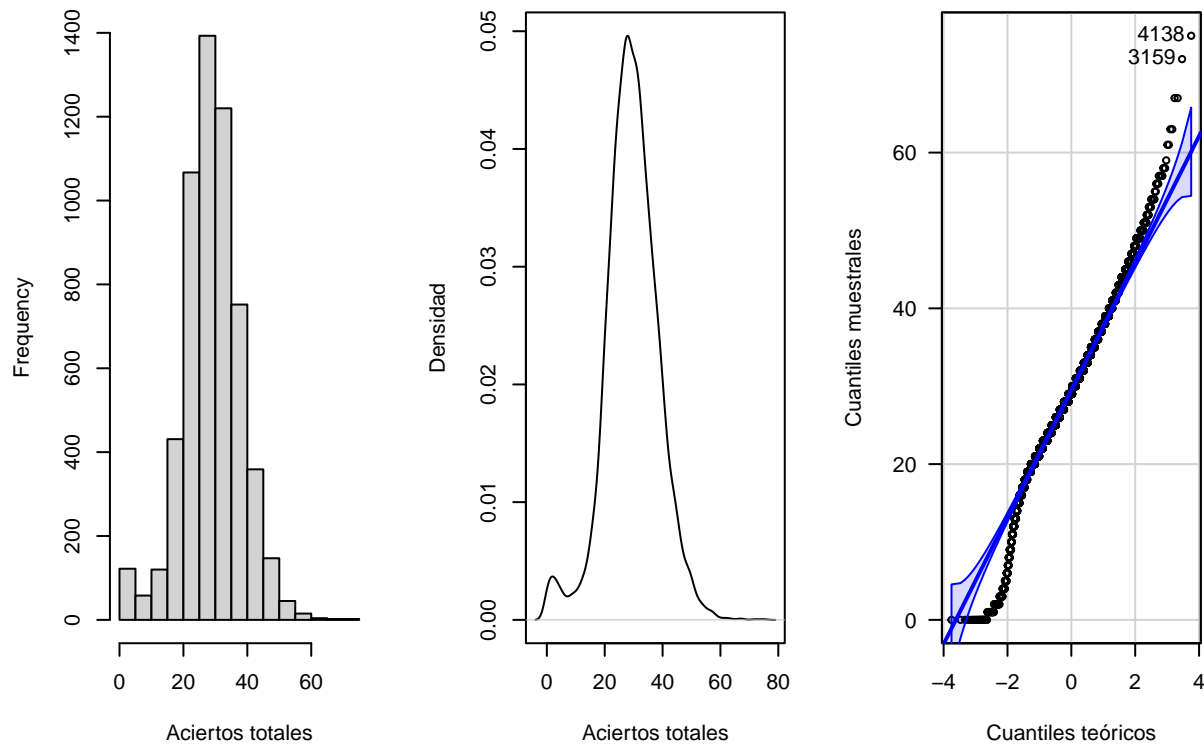
```
##      logit
```

```
##
```

```
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##   select
```

sity.default(x = datos\$`Aciertos 1



```
## [1] 4138 3159
```

Planteemos el estadístico:

$$D_n = \sup_x |\hat{F}_n(x) - F_0(x)|$$

Se calcula el estadístico por medio de la función `ks.test` y el valor p:

```
ktest = ks.test(datos$`Aciertos Totales`, "pnorm", media, desv)
```

```
## Warning in ks.test(datos$`Aciertos Totales`, "pnorm", media, desv): ties should
## not be present for the Kolmogorov-Smirnov test
```

```
ktest
```

```
##
## One-sample Kolmogorov-Smirnov test
##
## data:  datos$`Aciertos Totales`
## D = 0.055198, p-value = 1.332e-15
## alternative hypothesis: two-sided
```

Dado que la prueba de Anderson-Darling es más potente, se confirmará el resultado obtenido mediante la misma.

```
require(goftest)

## Loading required package: goftest

ad.test(datos$`Aciertos Totales`, "pnorm", media, desv)

##
## Anderson-Darling test of goodness-of-fit
## Null hypothesis: Normal distribution
## Parameters assumed to be fixed
##
## data:  datos$`Aciertos Totales`
## An = 21.904, p-value = 1.046e-07
```

Conclusión:

El valor del estadístico es $D = 0.055$ y un valor $p = 1.332e - 15$ en la prueba de Kolmogorov-Smirnov, para comprobar el resultado, se obtuvo un estadístico de $A_n = 21.904$ con un valor $p = 1.046e - 07$ en la prueba de Anderson-Darling, por lo que se rechaza rotundamente H_0 y concluimos que la variable *Total de Aciertos* no sigue una distribución normal. Otra evidencia que sustenta nuestra conclusión, son los valores *skew* y *kurtosis*, ya que si los mismos se encontraran en un rango entre -2 y 2, nos daría una temprana idea si es que nuestros datos están normalmente distribuidos, sin embargo, queda demostrado que en este caso no fue así.

Prueba de Hipótesis

Se quiere estimar la media poblacional de la cantidad de aciertos totales de los estudiantes que rindieron la prueba piloto EAES. Además, siendo la cantidad máxima 80 aciertos se quiere verificar si la media se encuentra al menos por encima o igual al 60% del máximo. Es decir: $\mu \geq 48$.

- Contraste de Hipotesis: $H_0 : \mu < 48$ vs. $H_1 : \mu \geq 48$.
- Media muestral: $\bar{x} = 29.41$.
- Estadístico de prueba: $Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = -154.1497$.
- Tamaño: $n = 5737$.
- Mediana: 29.
- Moda: 28.
- Sesgo: -5.536 .
- Alfa: $\alpha = 0.01\%$
- Intervalos de confianza: $[\bar{x} \pm Z_\alpha \sqrt{\frac{s}{n}}] = [29.09, 29.72]$.
- Valor-p: $p \approx 1$.

Entonces es posible decir que, la media poblacional de los aciertos totales es menor al 60% del número de aciertos posibles y el intervalo de la media al 99% está entre $29.09 \leq \mu \leq 29.72$, pero ya que son variables discretas resulta mejor concluir que en poco más del 99% la media se encuentra entre 29 y 30 aciertos.

Analisis de Contingencia

Al trabajar con variables categóricas, se puede resumir estos datos en una tabla de contingencia y la relación entre estas variables se puede verificar usando medidas de asociación. Se medira el grado de asociacion entre la variable **Forma** y **revRES**.

De manera general, la tabla de contingencia seria:

```
##      Cell Contents
## |-----|
## |              Count |
## |      Expected Values |
## |      Column Percent |
## |      Total Percent  |
## |-----|
##
## =====
##               reportes$revRES
## reportes$Forma      0      1      Total
## -----
## F001              19002      9078      28080
##                  17758.0    10322.0
##                  6.5%      5.4%
##                  4.1%      2.0%
## -----
## F002              19263      9617      28880
##                  18264.0    10616.0
##                  6.6%      5.7%
##                  4.2%      2.1%
## -----
## F003              18092      9668      27760
##                  17555.7    10204.3
##                  6.2%      5.7%
##                  3.9%      2.1%
## -----
## F004              17035      9445      26480
##                  16746.2    9733.8
##                  5.9%      5.6%
##                  3.7%      2.1%
## -----
## F005              21144      12056      33200
##                  20996.0    12204.0
##                  7.3%      7.1%
##                  4.6%      2.6%
## -----
## F006              20738      11662      32400
##                  20490.0    11910.0
##                  7.1%      6.9%
##                  4.5%      2.5%
## -----
## F007              20876      12724      33600
##                  21248.9    12351.1
##                  7.2%      7.5%
##                  4.5%      2.8%
```

```

## -----
## F008          18632      11768      30400
##              19225.2    11174.8
##              6.4%       7.0%
##              4.1%       2.6%
## -----
## F009          18127      9553       27680
##              17505.1    10174.9
##              6.2%       5.7%
##              3.9%       2.1%
## -----
## F010          16961      10879      27840
##              17606.3    10233.7
##              5.8%       6.4%
##              3.7%       2.4%
## -----
## F011          18237      10483      28720
##              18162.8    10557.2
##              6.3%       6.2%
##              4.0%       2.3%
## -----
## F012          16319      10961      27280
##              17252.1    10027.9
##              5.6%       6.5%
##              3.6%       2.4%
## -----
## F013          11815      8025       19840
##              12547.0    7293.0
##              4.1%       4.8%
##              2.6%       1.7%
## -----
## F014          12788      8492       21280
##              13457.7    7822.3
##              4.4%       5.0%
##              2.8%       1.8%
## -----
## F015          13646      7474       21120
##              13356.5    7763.5
##              4.7%       4.4%
##              3.0%       1.6%
## -----
## F016          13693      8707       22400
##              14166.0    8234.0
##              4.7%       5.2%
##              3.0%       1.9%
## -----
## F017          13933      8147       22080
##              13963.6    8116.4
##              4.8%       4.8%
##              3.0%       1.8%
## -----
## Total          290301    168739    459040
##              63.2%     36.8%
## =====

```


El V de Cramer es una medida del tamaño del efecto para la prueba chi-cuadrado de la independencia. En él se mide la forma en que están asociados dos campos categóricos. Expresado por:

$$V = \sqrt{\frac{\chi^2}{n(q-1)}}$$

Siendo q el mínimo entre el número de filas y de columnas.

```
## Loading required package: grid
```

```
assocstats(tabla)
```

```
##                X^2 df P(> X^2)
## Likelihood Ratio 1054.3 16      0
## Pearson          1052.1 16      0
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.048
## Cramer's V        : 0.048
```

Conclusion

Usando la V de Cramer 0.048 se puede concluir que el resultado es débil. Aunque el resultado es estadísticamente ligeramente significativo, los campos sólo están débilmente asociados (Forma~Respuesta).

Analisis Multivariante

Análisis de Correlación Canónico

Nuestro primer paso será definir ambas matrices que usaremos para el ACC, ya que recordemos que queremos estudiar la relación entre ambos pares de vectores U y V (base 1 y base 2).

```
## Loading required package: fda
```

```
## Loading required package: splines
```

```
## Loading required package: Matrix
```

```
## Loading required package: fda
```

```
## Loading required package: rainbow
```

```
## Loading required package: pcaPP
```

```
## Loading required package: RCurl
```

```
##
```

```
## Attaching package: 'fda'
```

```

## The following object is masked from 'package:graphics':
##
##      matplot

## Loading required package: fields

## Loading required package: spam

## Loading required package: dotCall64

## Spam version 2.7-0 (2021-06-25) is loaded.
## Type 'help( Spam)' or 'demo( spam)' for a short introduction
## and overview of this package.
## Help for individual functions is also obtained by adding the
## suffix '.spam' to the function name, e.g. 'help( chol.spam)'.

##
## Attaching package: 'spam'

## The following object is masked from 'package:Matrix':
##
##      det

## The following objects are masked from 'package:base':
##
##      backsolve, forwardsolve

## Loading required package: viridis

## Loading required package: viridisLite

## See https://github.com/NCAR/Fields for
## an extensive vignette, other supplements and source code

##
## Attaching package: 'fields'

## The following object is masked from 'package:psych':
##
##      describe

## Registered S3 method overwritten by 'GGally':
##      method from
##      +.gg      ggplot2

#Importamos los datos
base1 <- read_excel("Base 1.xlsx")
colnames(base1) = c("PROVINCIA", "MATEMATICA", "LITERATURA", "CCNN", "EESS")
head(base1, 10)

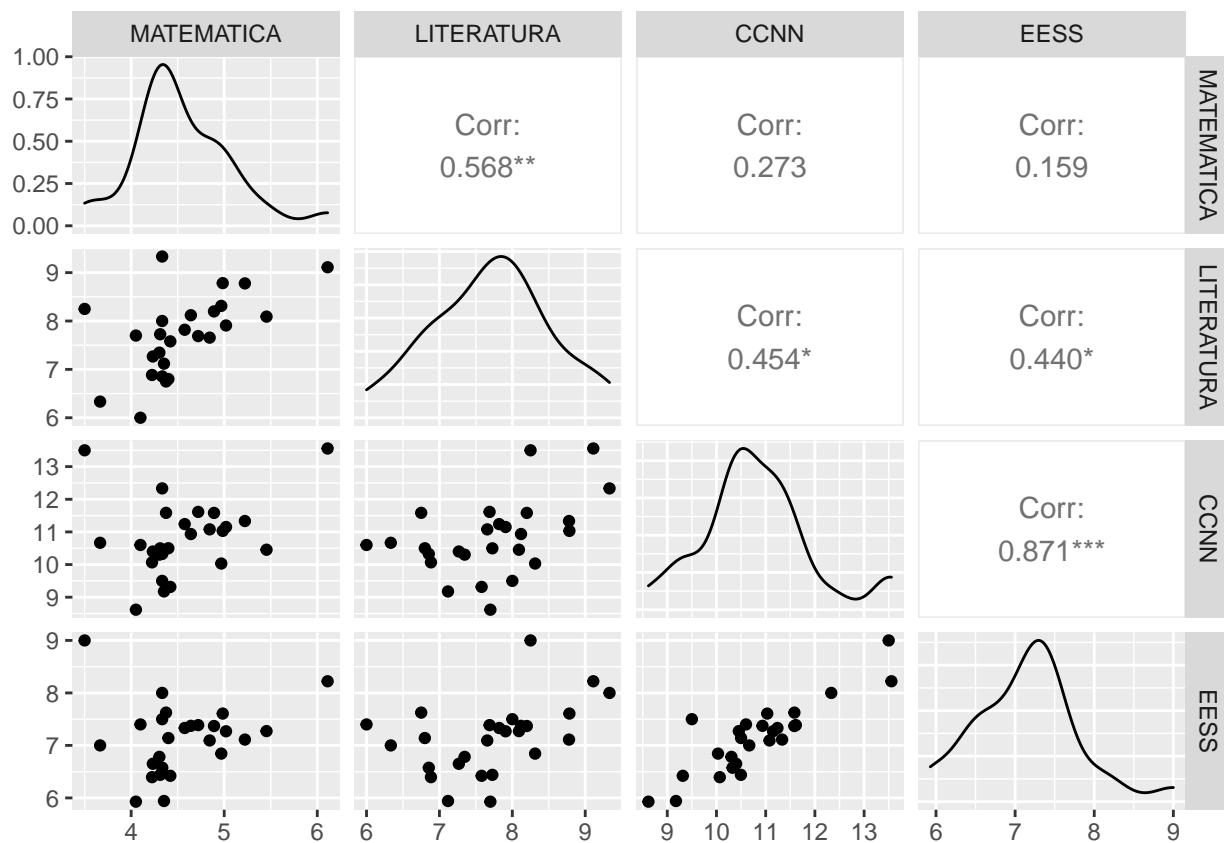
```

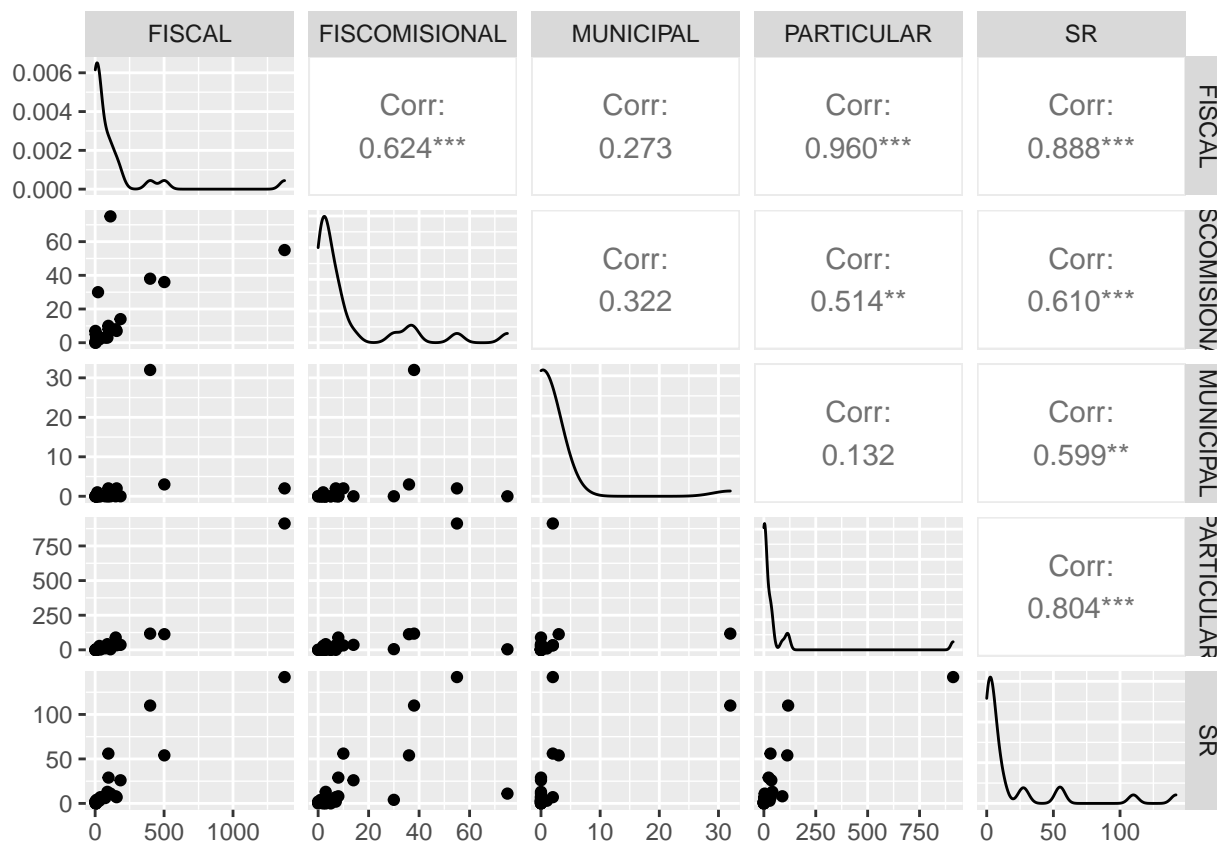
```
## # A tibble: 10 x 5
##   PROVINCIA      MATEMATICA LITERATURA  CCNN  EESS
##   <chr>          <dbl>      <dbl> <dbl> <dbl>
## 1 *Sin registro  4.05      7.70  8.62  5.93
## 2 AZUAY         4.33      9.33 12.3   8
## 3 BOLIVAR       4.4       6.8  10.5  7.14
## 4 CAÑAR        4.38      6.75 11.6   7.62
## 5 CARCHI        6.11      9.11 13.6   8.22
## 6 CHIMBORAZO    4.31      7.73 10.5   6.44
## 7 COTOPAXI     4.72      7.69 11.6   7.39
## 8 EL ORO        4.58      7.82 11.2   7.33
## 9 ESMERALDAS    4.23      7.27 10.4   6.65
## 10 GALAPAGOS    4.1       6    10.6  7.4
```

```
base2 <- read_excel("Base 2.xlsx",col_names=TRUE)
head(base2,10)
```

```
## # A tibble: 10 x 6
##   PROVINCIA      FISCAL FISCOMISIONAL MUNICIPAL PARTICULAR  SR
##   <chr>          <dbl>      <dbl>      <dbl>      <dbl> <dbl>
## 1 *Sin registro  96        10        2        32    56
## 2 AZUAY         12         1         0         2     3
## 3 BOLIVAR       38         3         0         2     7
## 4 CAÑAR        15         2         0         5     2
## 5 CARCHI         7         2         0         0     0
## 6 CHIMBORAZO    97         8         0        23    29
## 7 COTOPAXI     73         3         0        11     6
## 8 EL ORO       156         7         2        33     7
## 9 ESMERALDAS    21        30         0         5     4
## 10 GALAPAGOS     7         3         0         0     0
```

Visualizemos la correlación dentro de cada conjunto de variables:





En cuanto a las notas, se puede observar que hay una alta correlación entre las notas de Ciencias Sociales y Ciencias Naturales, y una muy baja entre Matemáticas y Ciencias Sociales. Al ser materias con syllabus tan diferentes, es esperado.

Por otro lado, en la otra base, tenemos que la correlación más alta está entre los estudiantes que vienen de colegios fiscales y particulares, mientras que la más baja está con los estudiantes de colegios municipales vs los de colegios particulares.

Una vez visto como se relacionan los datos, comencemos el ACC:

```
notas = base1[,2:5]
sostenimiento = base2[,2:6]
corr = matcor(notas,sostenimiento)
rownames(corr$XYcor) <- c("X1","X2","X3","X4","Y1","Y2","Y3","Y4","Y5")
kable( corr$XYcor , caption = "Matriz de correlación de los `datos`"
, align = c('l','c','c','c','c','c','c','c','c')
, col.names = c("X1","X2","X3","X4","Y1","Y2","Y3","Y4","Y5")
, row.names = TRUE
, digits = 3
, format.args = list( decimal.mark = "," )
)
```

Table 6: Matriz de correlación de los datos

	X1	X2	X3	X4	Y1	Y2	Y3	Y4	Y5
X1	1,000	0,568	0,273	0,159	-0,077	-0,008	0,138	-0,088	-0,055

	X1	X2	X3	X4	Y1	Y2	Y3	Y4	Y5
X2	0,568	1,000	0,454	0,440	-0,086	0,021	0,249	-0,094	0,035
X3	0,273	0,454	1,000	0,871	-0,131	-0,135	-0,005	-0,116	-0,209
X4	0,159	0,440	0,871	1,000	-0,175	-0,142	0,090	-0,147	-0,211
Y1	-0,077	-0,086	-0,131	-0,175	1,000	0,624	0,273	0,960	0,888
Y2	-0,008	0,021	-0,135	-0,142	0,624	1,000	0,322	0,514	0,610
Y3	0,138	0,249	-0,005	0,090	0,273	0,322	1,000	0,132	0,599
Y4	-0,088	-0,094	-0,116	-0,147	0,960	0,514	0,132	1,000	0,804
Y5	-0,055	0,035	-0,209	-0,211	0,888	0,610	0,599	0,804	1,000

```
acc = cc(notas,sostenimiento)
valores = as.vector(acc$cor)
corr_ca=diag(valores^0.5)
df = as.data.frame(corr_ca)
kable( df , caption = "Matriz de correlaciones canónicas"
, align = c('l','c','c','c','c','c','c','c','c')
, col.names = c("V1","V2","V3","V4")
, row.names = TRUE
, digits = 3
, format.args = list( decimal.mark = ","))
)
```

Table 7: Matriz de correlaciones canónicas

	V1	V2	V3	V4
1	0,758	0,000	0,000	0,000
2	0,000	0,676	0,000	0,000
3	0,000	0,000	0,297	0,000
4	0,000	0,000	0,000	0,186

Obteniendo las matrices de correlación entre U y X, y V y para hacer las debidas interpretaciones:

```
dfxu = as.data.frame(acc$scores$corr.X.xscores)
dfxu =t(dfxu)
rownames(dfxu) <- c("U1","U2","U3","U4")
dfyv = as.data.frame(acc$scores$corr.Y.yscores)
dfyv= t(dfyv)
rownames(dfyv) <- c("V1","V2","V3","V4")
kable( dfxu , caption = "Matriz de correlaciones canónicas"
, align = c('c','c','c','c','c')
, row.names = TRUE
, digits = 3
, format.args = list( decimal.mark = ","))
)
```

Table 8: Matriz de correlaciones canónicas

	MATEMATICA	LITERATURA	CCNN	EESS
U1	0,355	0,085	0,717	0,818

	MATEMATICA	LITERATURA	CCNN	EESS
U2	-0,295	-0,703	0,053	-0,276
U3	0,874	0,471	0,157	-0,159
U4	-0,153	0,527	0,677	0,479

```
kable( dfyv , caption = "Matriz de correlaciones canónicas"
, align = c('c','c','c','c','c')
, row.names = TRUE
, digits = 3
, format.args = list( decimal.mark = ",")
)
```

Table 9: Matriz de correlaciones canónicas

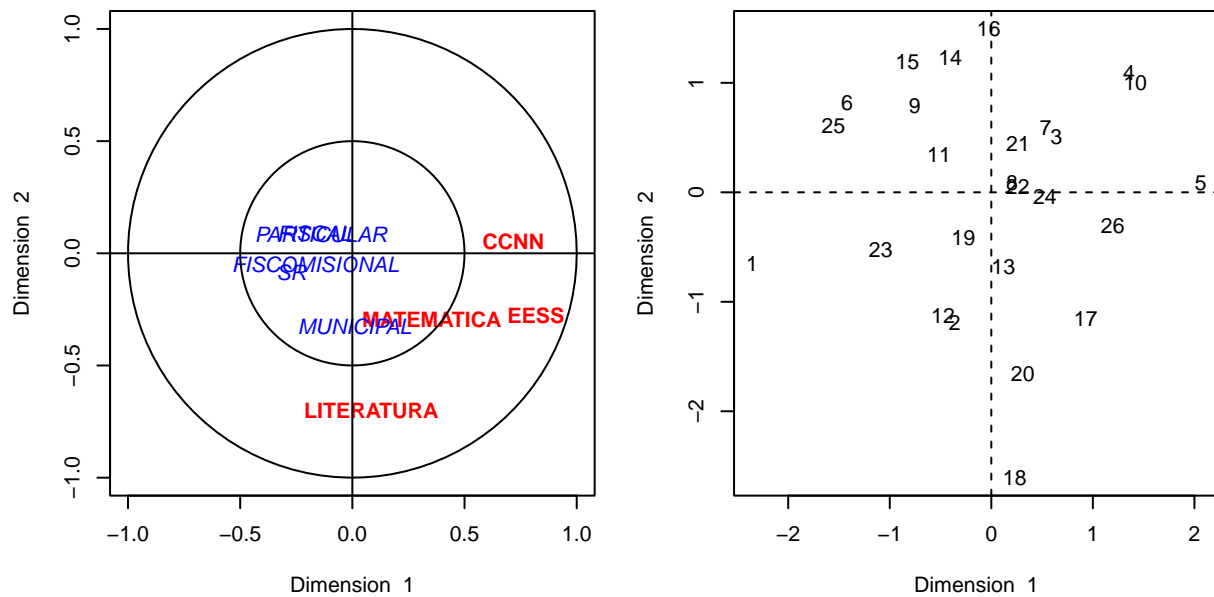
	FISCAL	FISCOMISIONAL	MUNICIPAL	PARTICULAR	SR
V1	-0,294	-0,287	0,018	-0,245	-0,474
V2	0,194	-0,104	-0,715	0,188	-0,185
V3	0,066	0,414	0,485	-0,210	0,188
V4	-0,675	-0,855	-0,100	-0,694	-0,496

En la matriz de correlación entre U y X se aprecian, entre otras, las siguientes correlaciones: U1 y *MATEMÁTICAS*: 0.355; U1 y *LITERATURA*: 0.085; U1 y *CCNN*: 0.717; U1 y *EESS*: 0.818. Así, U1 está altamente correlacionado, de manera directa, con las notas de Ciencias Naturales y Ciencias Sociales, pero correlacionado en menor magnitud con Matemáticas y Literatura.

Si analizamos el par U_1 y V_1 , podemos deducir que el promedio de notas en ambas asignaturas de ciencias es elevada para los estudiantes provenientes de un colegio municipal, pero el nivel de desempeño es relativamente menor en el resto de categorías de colegios

Si graficamos los resultados obtenidos:

```
plt.cc(acc,var.label=T)
```



Acerca del biplot generado, podemos decir que las materias representadas de mejor manera en la dimensión 1, ya que los mismos se encuentran fuera del primer círculo, por lo que los vectores tienen una norma más grande. Una forma más clara de representarlo sería mediante los vectores que salen desde el origen a las nuevas variables, por ejemplo, de ese análisis podemos notar que:

- Las nuevas variables de sostenimiento tienen una alta influencia en la dimensión 2.
- Los valores aberrante que encontramos en el gráfico 2 representan a las siguientes provincias: Carchi, Orellana y la zona sin registrar.
- Azuay, El Oro y Tungurahua, son las provincias con las notas más cercanas al centro definido y con unos aciertos de 5,7,11 y 7 para cada una de las materias respectivamente.

Analisis de Componentes Principales

Analizaremos el numero de componentes o factores a usar entre las variables de **Aciertos** en cada campo de conocimiento.

```
df1<-data.frame(componentes$eig,cumsum(componentes$eig),componentes$eig/4)

colnames(df1)<-c("Valor propio","Acumulado","Prop. acumulado")
rownames=rownames(df1)<-c("$\\lambda_1$","$\\lambda_2$","$\\lambda_3$","$\\lambda_4$")
kable(t(df1),caption="Valores propios desde la matriz de correlaciones",
      ,align=c('c','c','c','c')
      ,row.names=TRUE
      ,digits=3)
```



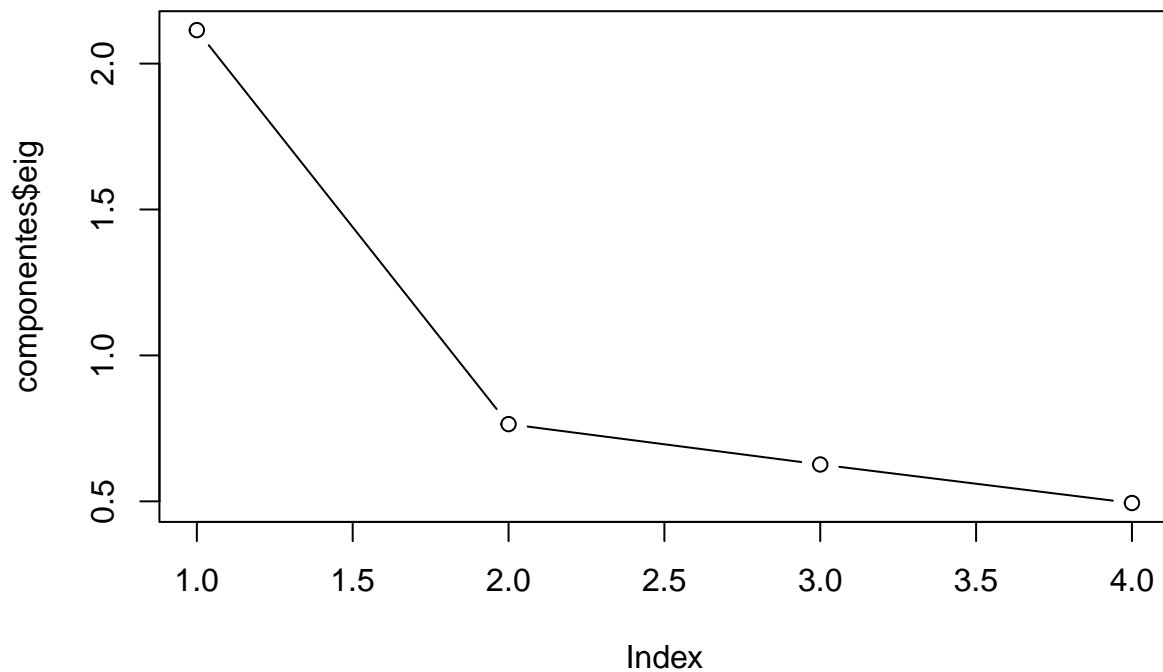
```
),format.args = list(decimal.mark=",")
)
```

Table 10: Valores propios desde la matriz de correlaciones

	λ_1	λ_2	λ_3	λ_4
Valor propio	2,115	0,765	0,626	0,494
Acumulado	2,115	2,879	3,506	4,000
Prop. acumulado	0,529	0,191	0,157	0,124

El 72% de la varianza se encuentra contenida en las componentes 1 y 2.

```
plot(componentes$eig,type = "b")#Grafico de valores propios
```



```
kable(componentes$c1,caption="Vectores Propios",digits=3,row.names=T)
```

Table 11: Vectores Propios

	CS1	CS2	CS3	CS4
Aciertos Matemáticas	-0.424	0.890	0.017	0.166
Aciertos Literatura	-0.500	-0.191	-0.828	-0.166

	CS1	CS2	CS3	CS4
Aciertos Ciencias Naturales	-0.538	-0.140	0.491	-0.671
Aciertos Ciencias Sociales	-0.530	-0.389	0.269	0.704

Entonces para el primer factor seria:

$$Y_1 = -0.42413X_1 - 0.50X_2 - 0.5377X_3 - 0.529995X_4$$

El segundo factor:

$$Y_2 = 0.89X_1 - 0.1912X_2 - 0.1405X_3 - 0.3893X_4$$

```
point_componentes<-cbind(dataTrab,componentes$li)
kable(head(point_componentes,n=100),caption="Valores propios desde la matriz de correlaciones"
, col.names = c("Codigo","Aciertos Matematica","Aciertos Literatura","Aciertos Ciencias Naturales")
, align=c('l','c','c','c','c','c','c','c','c','c')
, row.names=TRUE
, digits=1
, format.args = list(decimal.mark=",")
)
```

Table 12: Valores propios desde la matriz de correlaciones

	Codigo	Aciertos Matemat- ica	Aciertos Litera- tura	Aciertos Ciencias Naturales	Aciertos Sociales	Aciertos Totales	Prov	Eje1	Eje2	Eje3	Eje4
1	6563	5	5	3	0	13	09	2,5	1,5	-	-
2	3315	4	5	2	0	11	16	2,8	1,1	0,8	0,2
3	8140	5	5	10	4	24	09	0,9	0,8	1,0	0,1
4	3561	6	11	12	11	40	01	-	-	-	0,4
5	6050	5	9	11	6	31	01	1,8	0,1	0,3	0,6
6	4680	8	12	14	6	40	01	-	0,2	-	-
7	8070	3	7	10	11	31	01	0,2	1,2	0,4	0,3
8	7105	3	5	9	5	22	01	1,7	-	0,8	0,7
9	2829	2	6	9	10	27	01	-	-	0,4	1,0
10	4697	7	10	12	14	43	11	0,3	1,1	-	-
11	6490	4	12	16	11	43	01	1,2	-	0,3	-
12	5584	5	11	19	10	45	09	-	0,2	-	0,2
13	4405	6	13	21	11	51	01	0,4	-	0,5	0,9
								-	1,3	-	-
								2,3	0,1	0,2	1,5
								-	-	-	-
								2,1	1,2	0,1	0,2
								-	-	0,4	-
								2,3	0,7	-	0,8
								-	-	0,2	-
								3,3	0,6	-	0,9

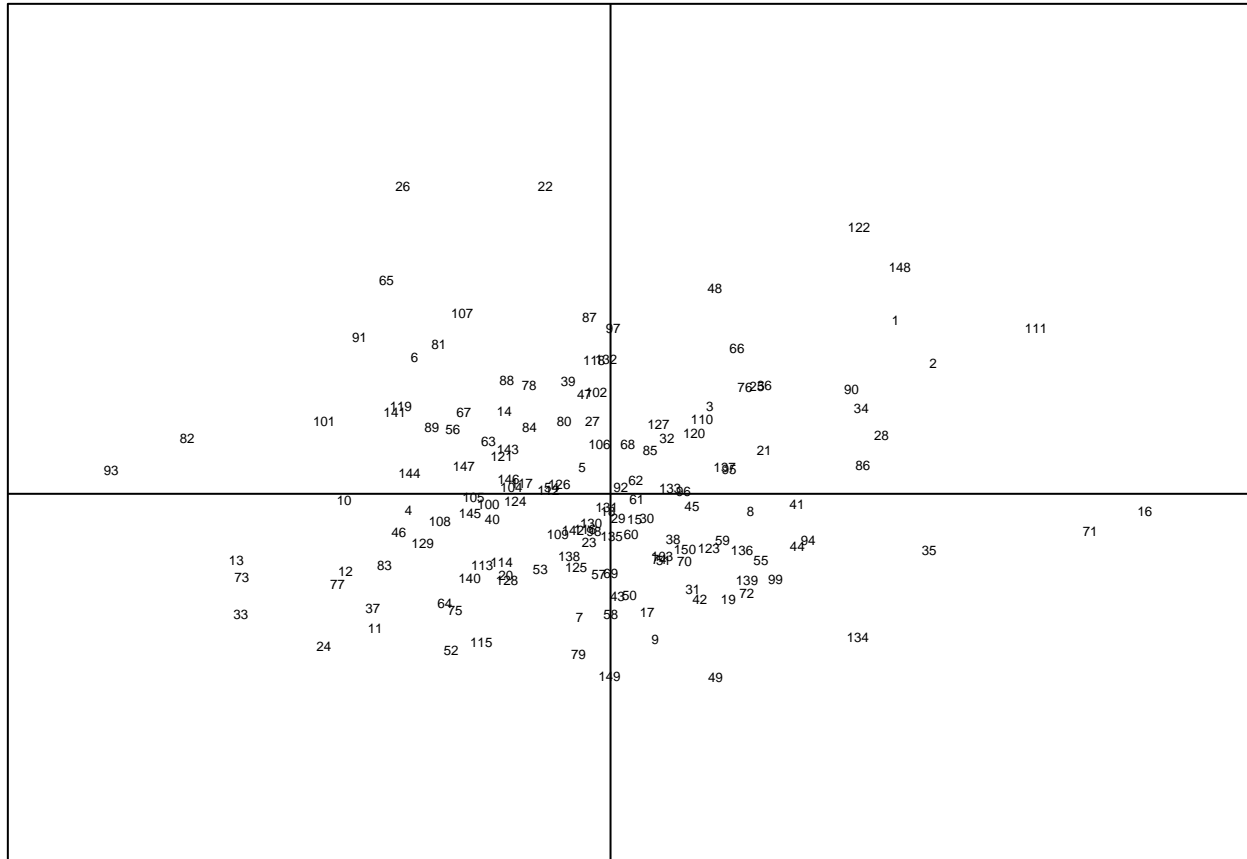
	Codigo	Aciertos Matemat- ica	Aciertos Liter- atura	Aciertos Ciencias Naturales	Aciertos Sociales	Aciertos Totales	Prov	Eje1	Eje2	Eje3	Eje4
14	1484	7	10	8	9	34	01	- 0,9	0,7	- 0,7	1,0
15	3468	4	6	10	8	28	17	0,2	- 0,2	0,5	0,4
16	7108	0	0	0	0	0	17	4,7	- 0,2	0,1	0,2
17	3357	2	10	10	6	28	01	0,3	- 1,0	- 0,8	- 0,4
18	4620	4	8	12	6	30	01	0,0	- 0,2	0,0	- 0,5
19	7873	2	5	8	8	23	12	1,0	- 0,9	0,5	0,6
20	4613	4	12	10	9	35	17	- 0,9	- 0,7	- 1,0	0,3
21	8084	4	3	9	5	21	02	1,3	0,4	0,9	0,0
22	376	10	7	11	4	32	02	- 0,6	2,7	0,0	- 0,3
23	1717	4	11	7	8	30	18	- 0,2	- 0,4	- 1,2	0,6
24	4227	4	14	17	11	46	17	- 2,5	- 1,3	- 0,5	- 0,5
25	5772	5	6	7	3	21	02	1,3	0,9	- 0,3	- 0,2
26	625	11	10	14	5	40	*Sin reg- istro	- 1,8	2,7	- 0,3	- 0,6
27	7701	6	5	11	8	30	12	- 0,2	0,6	0,9	0,4
28	5688	3	3	8	1	15	12	2,4	0,5	0,4	- 0,8
29	6186	4	8	10	7	29	02	0,1	- 0,2	- 0,2	0,0
30	4413	4	7	8	8	27	17	0,3	- 0,2	- 0,1	0,7
31	5741	2	5	13	6	26	02	0,7	- 0,8	0,9	- 0,7
32	5037	5	6	9	6	26	02	0,5	0,5	0,2	0,2
33	5083	5	14	21	11	51	02	- 3,2	- 1,1	- 0,1	- 1,1
34	5112	4	4	4	3	15	02	2,2	0,7	- 0,2	0,3
35	5016	1	4	4	3	12	02	2,8	- 0,5	- 0,2	0,1
36	7936	5	2	10	4	21	02	1,4	1,0	1,2	- 0,3
37	5252	4	11	20	9	44	02	- 2,1	- 1,0	0,4	- 1,3
38	5209	3	4	14	6	27	02	0,5	- 0,4	1,3	- 0,7

	Codigo	Aciertos Matemat- ica	Aciertos Liter- atura	Aciertos Ciencias Naturales	Aciertos Sociales	Aciertos Totales	Prov	Eje1	Eje2	Eje3	Eje4
39	6144	7	4	11	9	31	02	- 0,4	1,0	1,2	0,8
40	7587	5	9	13	9	36	02	- 1,0	- 0,2	0,1	0,0
41	4350	3	3	7	6	19	02	1,6	- 0,1	0,7	0,5
42	4342	2	6	10	7	25	18	0,8	- 0,9	0,3	0,0
43	5323	3	6	10	10	29	02	0,1	- 0,9	0,6	0,8
44	8266	2	2	11	5	20	02	1,6	- 0,5	1,3	- 0,4
45	7417	4	7	5	8	24	02	0,7	- 0,1	- 0,4	1,2
46	3405	6	11	10	13	40	02	- 1,9	- 0,3	- 0,4	1,4
47	3455	6	9	12	4	31	*Sin reg- istro	- 0,2	0,9	- 0,4	- 0,9
48	7372	7	5	8	3	23	02	0,9	1,8	0,1	- 0,2
49	3625	1	2	10	11	24	02	0,9	- 1,6	1,7	1,1
50	7375	3	7	8	10	28	18	0,2	- 0,9	0,1	1,0
51	5839	3	4	12	8	27	02	0,5	- 0,6	1,2	0,1
52	6549	3	11	15	10	39	02	- 1,4	- 1,4	- 0,1	- 0,3
53	2899	4	9	10	10	33	17	- 0,6	- 0,7	- 0,2	0,7
54	5964	5	6	14	8	33	02	- 0,5	0,1	0,9	- 0,2
55	2929	2	4	11	5	22	06	1,3	- 0,6	0,8	- 0,5
56	707	7	12	9	9	37	02	- 1,4	0,6	- 1,1	0,7
57	3656	3	7	14	7	31	02	- 0,1	- 0,7	0,6	- 0,6
58	6674	3	5	9	12	29	02	0,0	- 1,1	0,9	1,5
59	1713	3	6	7	7	23	17	1,0	- 0,4	0,0	0,6
60	5255	4	4	10	10	28	02	0,2	- 0,4	1,2	1,0
61	1924	4	10	9	5	28	02	0,2	0,0	- 1,0	- 0,4
62	4329	5	13	0	8	26	02	0,2	0,1	- 2,6	1,7

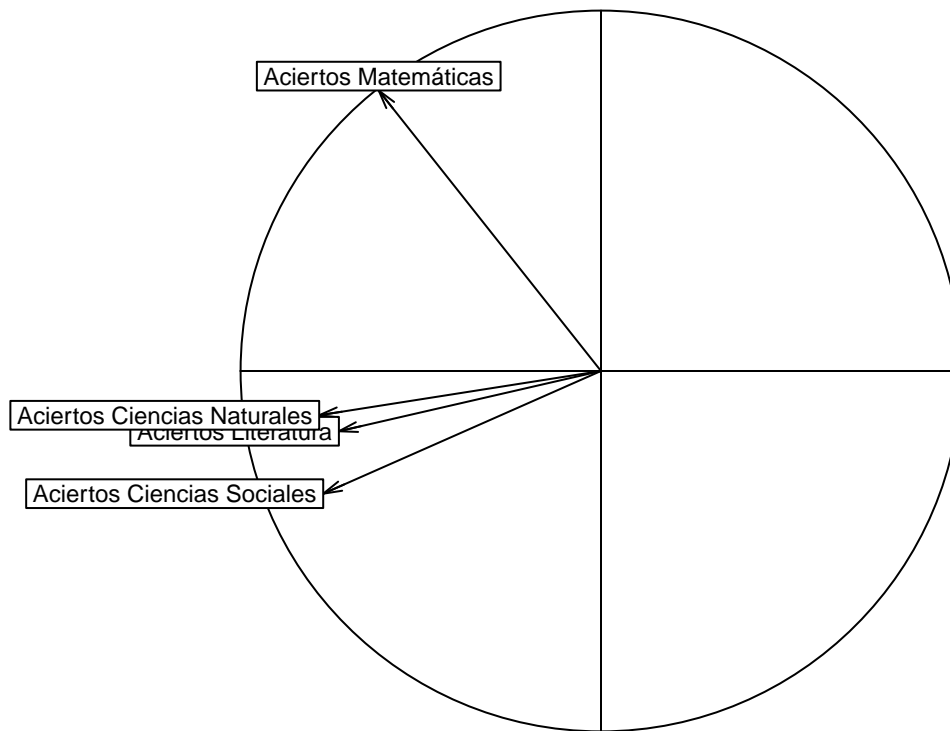
	Codigo	Aciertos Matemat- ica	Aciertos Liter- atura	Aciertos Ciencias Naturales	Aciertos Sociales	Aciertos Totales	Prov	Eje1	Eje2	Eje3	Eje4
63	5181	6	6	18	7	37	03	- 1,1	0,5	1,3	- 1,0
64	7863	4	9	15	11	39	90	- 1,5	- 1,0	0,5	0,1
65	7997	10	5	16	10	41	03	- 2,0	1,9	1,7	0,4
66	6293	6	2	9	5	22	03	1,1	1,3	1,1	0,2
67	6501	7	10	12	8	37	09	- 1,3	0,7	- 0,3	0,1
68	7963	5	4	14	6	29	03	0,2	0,4	1,3	- 0,6
69	6729	3	8	12	7	30	09	0,0	- 0,7	0,1	- 0,4
70	3316	3	7	7	8	25	03	0,6	- 0,6	- 0,2	0,7
71	3378	0	3	0	0	3	03	4,2	- 0,3	- 0,7	0,0
72	7431	1	6	15	2	24	03	1,2	- 0,9	0,5	- 2,1
73	7910	6	13	18	13	50	03	- 3,2	- 0,7	0,0	0,0
74	6067	3	7	10	7	27	03	0,4	- 0,6	0,1	0,0
75	5988	4	9	13	12	38	03	- 1,4	- 1,0	0,4	0,7
76	2809	5	5	9	3	22	04	1,2	0,9	0,2	- 0,5
77	7194	5	12	17	11	45	04	- 2,4	- 0,8	0,0	- 0,3
78	588	7	10	9	7	33	04	- 0,7	0,9	- 0,8	0,3
79	7285	2	8	13	9	32	04	- 0,3	- 1,4	0,4	- 0,1
80	1373	6	6	13	7	32	04	- 0,4	0,6	0,7	- 0,2
81	7376	8	9	16	6	39	04	- 1,5	1,3	0,2	- 0,9
82	7163	9	12	17	14	52	04	- 3,7	0,5	0,3	0,7
83	4450	5	13	14	10	42	17	- 2,0	- 0,6	- 0,7	- 0,1
84	2194	6	9	13	6	34	17	- 0,7	0,6	- 0,1	- 0,6
85	1034	5	5	10	7	27	13	0,3	0,4	0,6	0,3
86	771	3	6	3	3	15	*Sin reg- istro	2,2	0,2	- 0,8	0,3
87	2935	8	6	7	8	29	05	- 0,2	1,5	0,1	1,2

	Codigo	Aciertos Matemat- ica	Aciertos Liter- atura	Aciertos Ciencias Naturales	Aciertos Sociales	Aciertos Totales	Prov	Eje1	Eje2	Eje3	Eje4
88	3539	7	9	13	6	35	17	- 0,9	1,0	- 0,1	- 0,5
89	3351	7	9	14	9	39	05	- 1,6	0,6	0,3	0,0
90	3619	4	7	5	0	16	05	2,1	0,9	- 1,1	- 0,7
91	7440	9	8	17	9	43	05	- 2,2	1,4	0,9	- 0,3
92	7220	4	7	15	4	30	05	0,1	0,1	0,4	- 1,4
93	326	9	9	23	16	57	05	- 4,4	0,2	2,0	0,3
94	3430	2	5	8	4	19	17	1,7	- 0,4	0,1	- 0,3
95	7197	4	3	10	6	23	05	1,0	0,2	1,1	0,1
96	208	4	8	7	6	25	*Sin reg- istro	0,6	0,0	- 0,6	0,3
97	644	7	7	11	4	29	17	0,0	1,4	0,0	- 0,5
98	795	4	11	8	7	30	*Sin reg- istro	- 0,1	- 0,3	- 1,2	0,2
99	3278	2	6	5	7	20	05	1,4	- 0,8	- 0,3	0,8
100	371	5	15	10	6	36	*Sin reg- istro	- 1,1	- 0,1	- 2,1	- 0,5

```
s.label(head(componentes$li,n=150),grid = F,boxes = F,clabel = 0.45)#Grafico sobre el primer plano fact
```



```
s.corcircle(componentes$co,grid = F,clabel = 0.75)
```



Conclusion (Observaciones):

Las observaciones 26, 22, 65 o 122 tienen valores altos en el 2^{do} factor que como se desglosó los pesos, este factor está influenciado de mayor manera por las variables de Matemáticas y Ciencias Sociales. Lo que es contraintuitivo a lo que dicen las opiniones de los expertos. Así que sería lógico revisar las preguntas formuladas en los cuestionarios de tal campo de conocimiento.

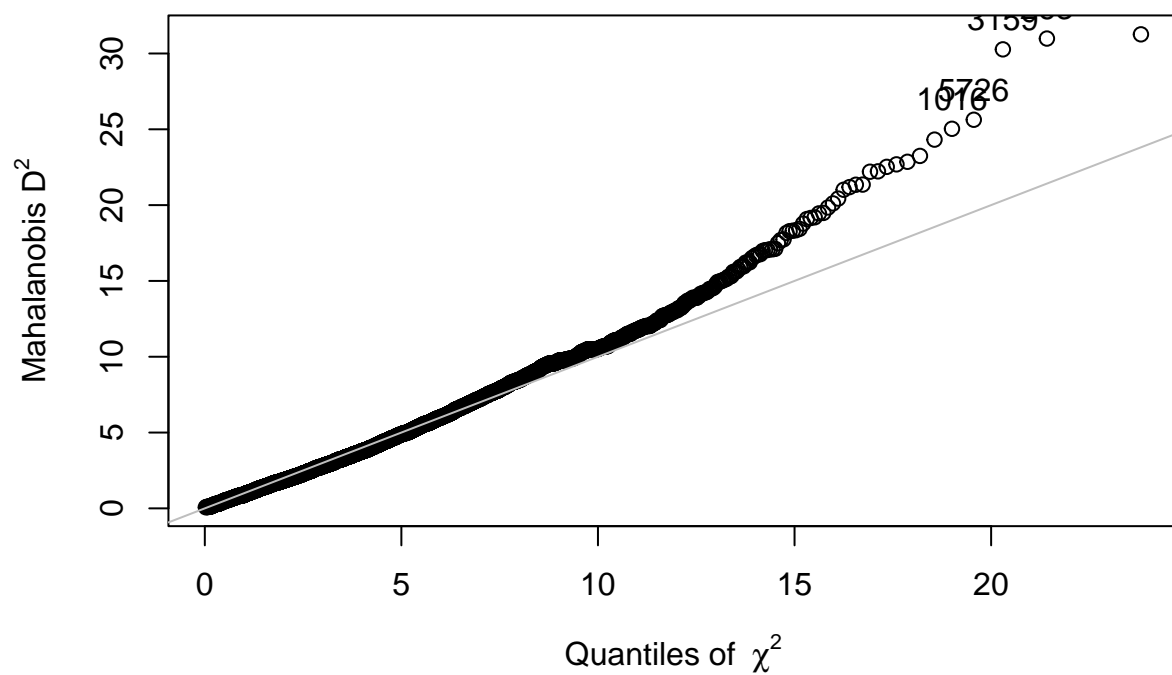
Analisis Factorial

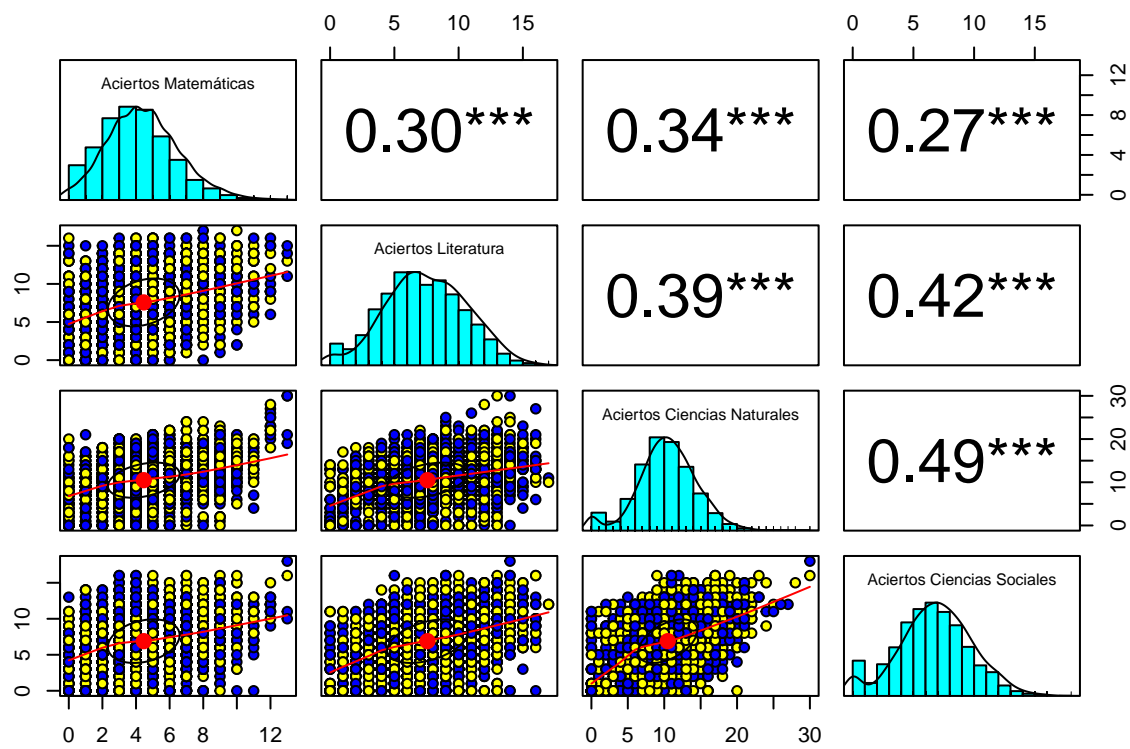
El análisis factorial se utiliza principalmente para la reducción de datos o la detección de estructuras. El propósito de la detección de estructuras es examinar las relaciones subyacentes (o latentes) entre las variables.

```
##
## Attaching package: 'gridExtra'

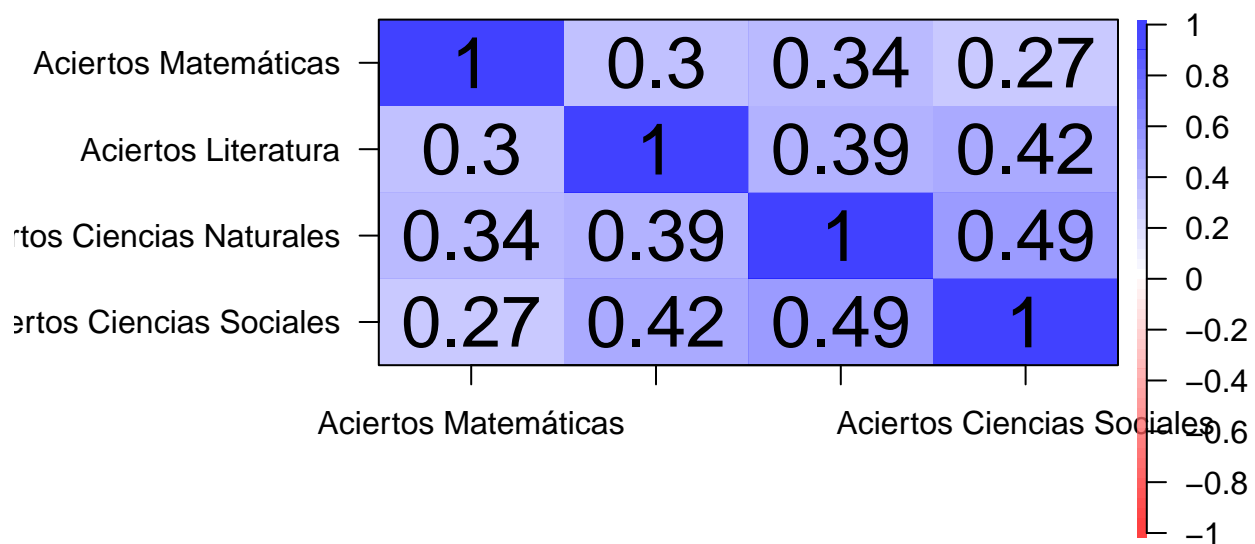
## The following object is masked from 'package:dplyr':
##
## combine
```

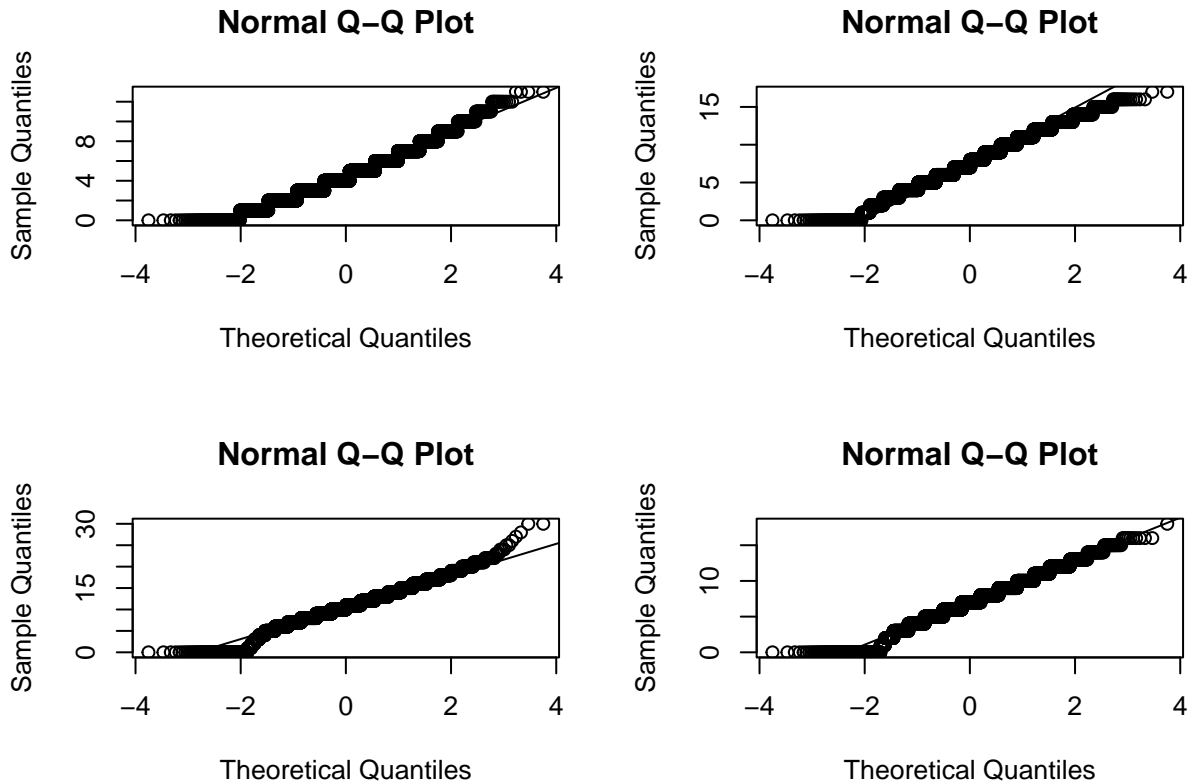

Q-Q plot of Mahalanobis D^2 vs. quantiles of χ^2_{nvar}





```
cor.plot(cor(dataTrab[,2:5]),numbers = T)
```





```
##
## Attaching package: 'nortest'

## The following objects are masked from 'package:gofstest':
##
##   ad.test, cvm.test

## Warning in cvm.test(dataTrab$'Aciertos Matemáticas'): p-value is smaller than
## 7.37e-10, cannot be computed more accurately
```

Table 13: Test de Normalidad Univariante

	Anderson.Darling	Cramer.von.Mises	Lilliefors..Kolmogorov.Smirnov.	Pearson.chi.square
Estadistico	60.18847	11.23793	0.1141524	43807.4
Valor-p	0.00000	0.00000	0.0000000	0.0

```
ad2=ad.test(dataTrab$`Aciertos Literatura`)
cvm2=cvm.test(dataTrab$`Aciertos Literatura`)
```

```
## Warning in cvm.test(dataTrab$'Aciertos Literatura'): p-value is smaller than
## 7.37e-10, cannot be computed more accurately
```

```

lil2=lillie.test(dataTrab$`Aciertos Literatura`)
pear2=pearson.test(dataTrab$`Aciertos Literatura`)
norm2<-data.frame("Anderson-Darling"=c(ad2$statistic,ad2$p.value),
                  "Cramer-von Mises"=c(cvm2$statistic,cvm2$p.value),
                  "Lilliefors (Kolmogorov-Smirnov)"=c(lil2$statistic,lil2$p.value),
                  "Pearson chi-square"=c(pear2$statistic,pear2$p.value),
                  row.names=c("Estadistico","Valor-p"))
kable(norm2,caption = "Test de Normalidad Univariante")

```

Table 14: Test de Normalidad Univariante

	Anderson.Darling	Cramer.von.Mises	Lilliefors..Kolmogorov.Smirnov.	Pearson.chi.square
Estadistico	26.19024	4.775842	0.0756372	26985.07
Valor-p	0.00000	0.000000	0.0000000	0.00

```

ad3=ad.test(dataTrab$`Aciertos Ciencias Naturales`)
cvm3=cvm.test(dataTrab$`Aciertos Ciencias Naturales`)

```

```

## Warning in cvm.test(dataTrab$`Aciertos Ciencias Naturales`): p-value is smaller
## than 7.37e-10, cannot be computed more accurately

```

```

lil3=lillie.test(dataTrab$`Aciertos Ciencias Naturales`)
pear3=pearson.test(dataTrab$`Aciertos Ciencias Naturales`)
norm3<-data.frame("Anderson-Darling"=c(ad3$statistic,ad3$p.value),
                  "Cramer-von Mises"=c(cvm3$statistic,cvm3$p.value),
                  "Lilliefors (Kolmogorov-Smirnov)"=c(lil3$statistic,lil3$p.value),
                  "Pearson chi-square"=c(pear3$statistic,pear3$p.value),
                  row.names=c("Estadistico","Valor-p"))
kable(norm3,caption = "Test de Normalidad Univariante")

```

Table 15: Test de Normalidad Univariante

	Anderson.Darling	Cramer.von.Mises	Lilliefors..Kolmogorov.Smirnov.	Pearson.chi.square
Estadistico	29.73758	4.657694	0.0725304	22133.95
Valor-p	0.00000	0.000000	0.0000000	0.00

```

ad4=ad.test(dataTrab$`Aciertos Ciencias Sociales`)
cvm4=cvm.test(dataTrab$`Aciertos Ciencias Sociales`)

```

```

## Warning in cvm.test(dataTrab$`Aciertos Ciencias Sociales`): p-value is smaller
## than 7.37e-10, cannot be computed more accurately

```

```

lil4=lillie.test(dataTrab$`Aciertos Ciencias Sociales`)
pear4=pearson.test(dataTrab$`Aciertos Ciencias Sociales`)
norm4<-data.frame("Anderson-Darling"=c(ad4$statistic,ad4$p.value),
                  "Cramer-von Mises"=c(cvm4$statistic,cvm4$p.value),
                  "Lilliefors (Kolmogorov-Smirnov)"=c(lil4$statistic,lil4$p.value),
                  "Pearson chi-square"=c(pear4$statistic,pear4$p.value),
                  row.names=c("Estadistico","Valor-p"))
kable(norm4,caption = "Test de Normalidad Univariante")

```

Table 16: Test de Normalidad Univariante

	Anderson.Darling	Cramer.von.Mises	Lilliefors..Kolmogorov.Smironov.	Pearson.chi.square
Estadistico	33.50121	5.687869	0.0778773	29945.72
Valor-p	0.00000	0.000000	0.0000000	0.00

```
# Test de Normalidad Multivariante
mvdemisdatos<-mvn(dataTrab[,2:5],mvnTest = "mardia")

mvdemisdatos$univariateNormality
```

```
##          Test          Variable Statistic    p value Normality
## 1 Anderson-Darling  Aciertos Matemáticas    60.1885 <0.001      NO
## 2 Anderson-Darling  Aciertos Literatura      26.1902 <0.001      NO
## 3 Anderson-Darling  Aciertos Ciencias Naturales 29.7376 <0.001      NO
## 4 Anderson-Darling  Aciertos Ciencias Sociales 33.5012 <0.001      NO
```

```
mvdemisdatos$multivariateNormality
```

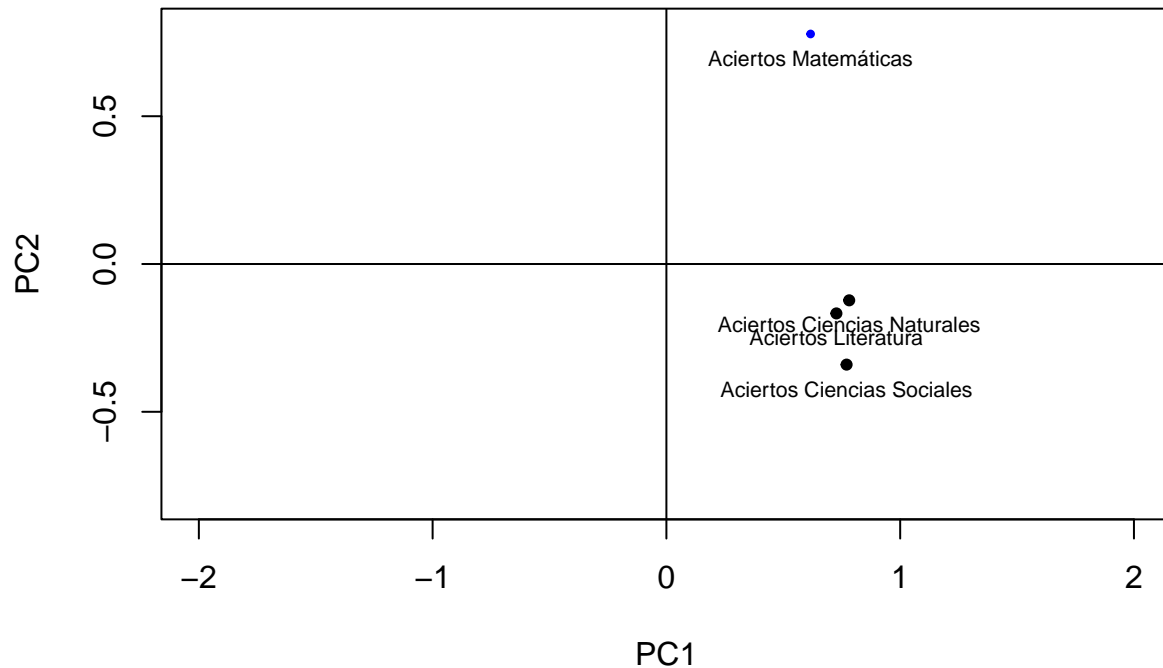
```
##          Test          Statistic          p value Result
## 1 Mardia Skewness 385.987258495808 1.63938719654121e-69    NO
## 2 Mardia Kurtosis 12.4878691773246          0          NO
## 3          MVN          <NA>          <NA>          NO
```

```
R=cor(dataTrab[,2:5])
fit.pca=principal(R,nfactors = 2,rotate = "none")
fit.pca
```

```
## Principal Components Analysis
## Call: principal(r = R, nfactors = 2, rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##          PC1  PC2  h2  u2 com
## Aciertos Matemáticas    0.62  0.78 0.99 0.014 1.9
## Aciertos Literatura      0.73 -0.17 0.56 0.443 1.1
## Aciertos Ciencias Naturales 0.78 -0.12 0.63 0.373 1.0
## Aciertos Ciencias Sociales 0.77 -0.34 0.71 0.290 1.4
##
##          PC1  PC2
## SS loadings    2.11 0.76
## Proportion Var    0.53 0.19
## Cumulative Var    0.53 0.72
## Proportion Explained 0.73 0.27
## Cumulative Proportion 0.73 1.00
##
## Mean item complexity = 1.4
## Test of the hypothesis that 2 components are sufficient.
##
## The root mean square of the residuals (RMSR) is 0.13
##
## Fit based upon off diagonal values = 0.87
```

```
plot(fit.pca,labels=row.names(R),cex=.7,xlim=c(-2,2),ylim=c(-0.8,0.8))
```

Principal Component Analysis



```
det(R)
```

```
## [1] 0.5006291
```

Un determinante muy bajo (cercano a cero) indicará altas intercorrelaciones entre las variables. Si el determinante llegase a cero indicaría la existencia de relaciones lineales perfectas entre las variables explicativas y por tanto presencia de multicolinealidad perfecta en el modelo.

```
KMO(dataTrab[,2:5])
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = dataTrab[, 2:5])
## Overall MSA = 0.73
## MSA for each item =
##      Aciertos Matemáticas      Aciertos Literatura
##                0.79                0.76
## Aciertos Ciencias Naturales Aciertos Ciencias Sociales
##                0.70                0.70
```

Interpretación del KMO: La medida de adecuación de la muestra *MSA* o *KMO* (*Kaiser-Meyer-Olkin*) contrasta si las correlaciones parciales entre las variables son suficientemente pequeñas. El estadístico KMO varía entre 0 y 1. Los valores altos (cerca de 1) generalmente indican que un análisis factorial puede ser útil con sus datos.

```
bartlett.test(dataTrab[,2:5])
```

```
##  
## Bartlett test of homogeneity of variances  
##  
## data: dataTrab[, 2:5]  
## Bartlett's K-squared = 2257.7, df = 3, p-value < 2.2e-16
```

Prueba de esfericidad de Bartlett: Prueba la hipótesis de que su matriz de correlación es una matriz de identidad, lo que indicaría que sus variables no están relacionadas y, por lo tanto, no son adecuadas para la detección de estructuras. Los valores pequeños (menos de 0.05) del nivel de significancia indican que un análisis factorial puede ser útil con sus datos.

```
fit.pa.none=fa(R,nfactores=2,fm="pa",rotate="none",n.obs=500)  
fit.pa.rot.vari= fa(R,nfactores=2,fm="pa",rotate="varimax",n.obs=500)  
fit.pa.rot.pro=fa(R,nfactores=2,fm="pa",rotate="quartimax",n.obs=500)  
  
fit.pa.none$uniquenesses
```

```
##          Aciertos Matemáticas          Aciertos Literatura  
##                0.7929374                0.6446052  
## Aciertos Ciencias Naturales Aciertos Ciencias Sociales  
##                0.5074053                0.5337916
```

```
fit.ml.none=fa(R,nfactors = 2,fm="ml",rotate = "none",n.obs = 500)  
fit.ml.rot.vari=fa(R,nfactors = 2,fm="ml",rotate = "varimax",n.obs = 500)  
fit.ml.rot.pro=fa(R,nfactors = 2,fm="ml",rotate = "quartimax",n.obs = 500)  
  
fit.ml.none$uniquenesses
```

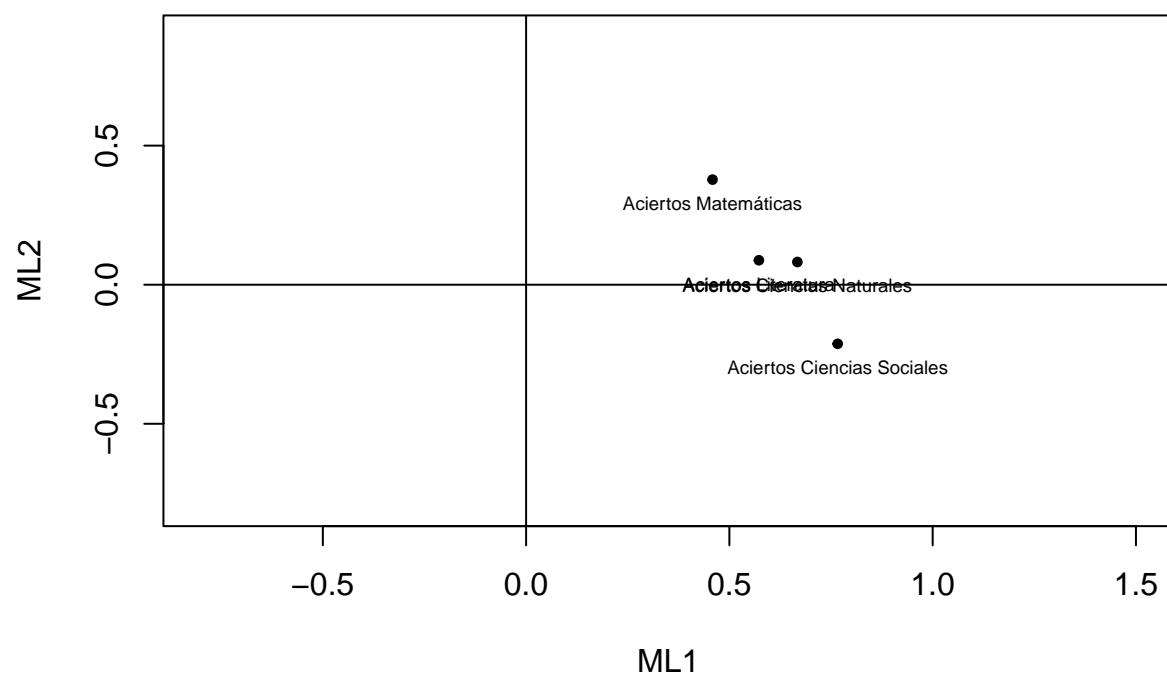
```
##          Aciertos Matemáticas          Aciertos Literatura  
##                0.6469717                0.6643881  
## Aciertos Ciencias Naturales Aciertos Ciencias Sociales  
##                0.5480667                0.3675426
```

```
fit.om.none=fa(R,nfactors = 2,fm="old.min",rotate = "none",n.obs = 500)  
fit.om.rot.vari=fa(R,nfactors = 2,fm="old.min",rotate = "varimax",n.obs = 500)  
fit.om.rot.pro=fa(R,nfactors = 2,fm="old.min",rotate = "quartimax",n.obs = 500)  
  
fit.om.none$uniquenesses
```

```
##          Aciertos Matemáticas          Aciertos Literatura  
##                0.1461871                0.6677845  
## Aciertos Ciencias Naturales Aciertos Ciencias Sociales  
##                0.5437111                0.4506423
```

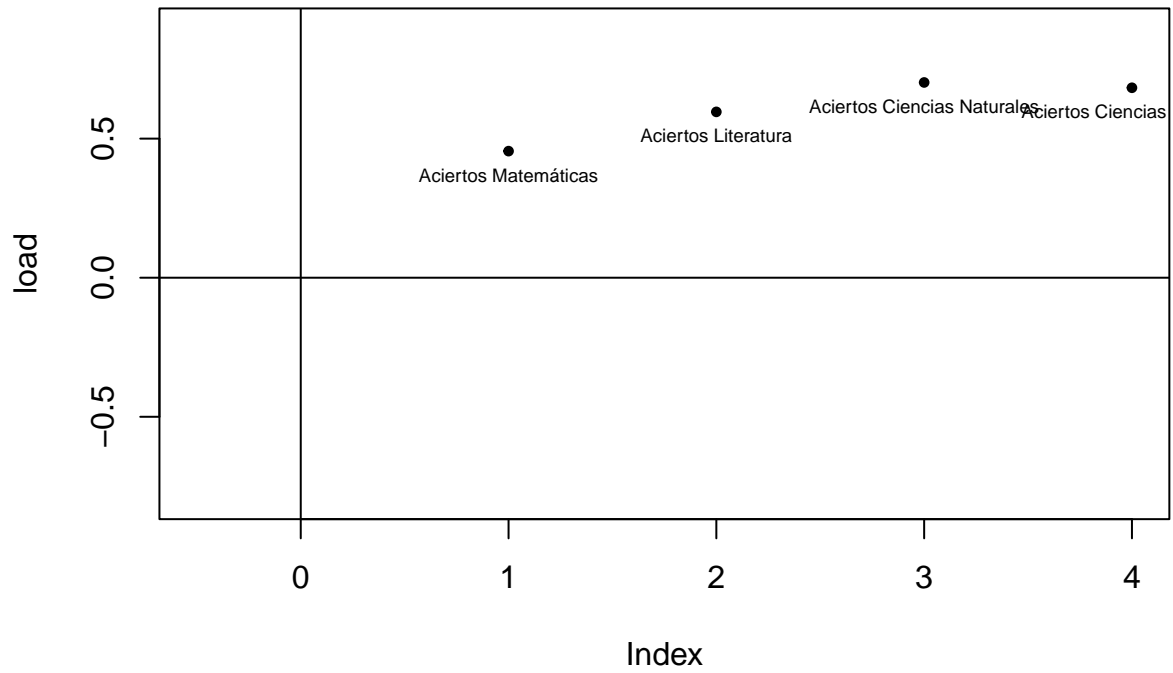
```
plot(fit.ml.none,labels = row.names(R),cex=0.6,xlim=c(-0.8,1.5),ylim=c(-0.8,0.9))
```


Factor Analysis



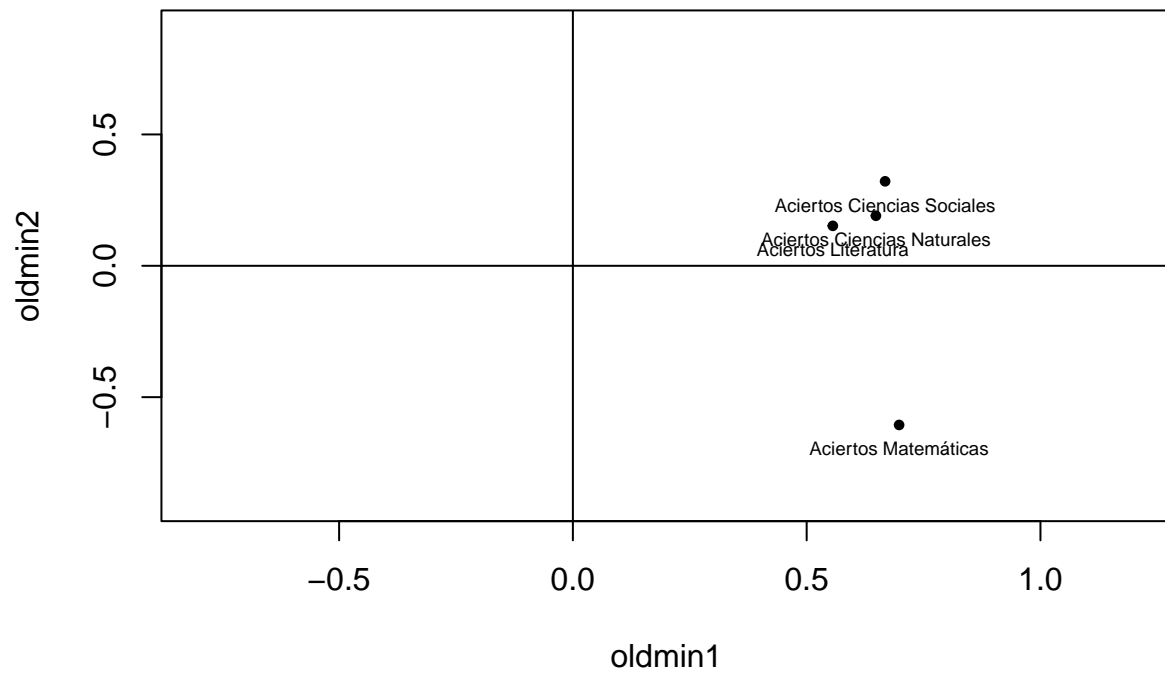
```
plot(fit.pa.none, labels = row.names(R), cex=0.6, xlim=c(-0.5, 4), ylim=c(-0.8, 0.9))
```

Factor Analysis



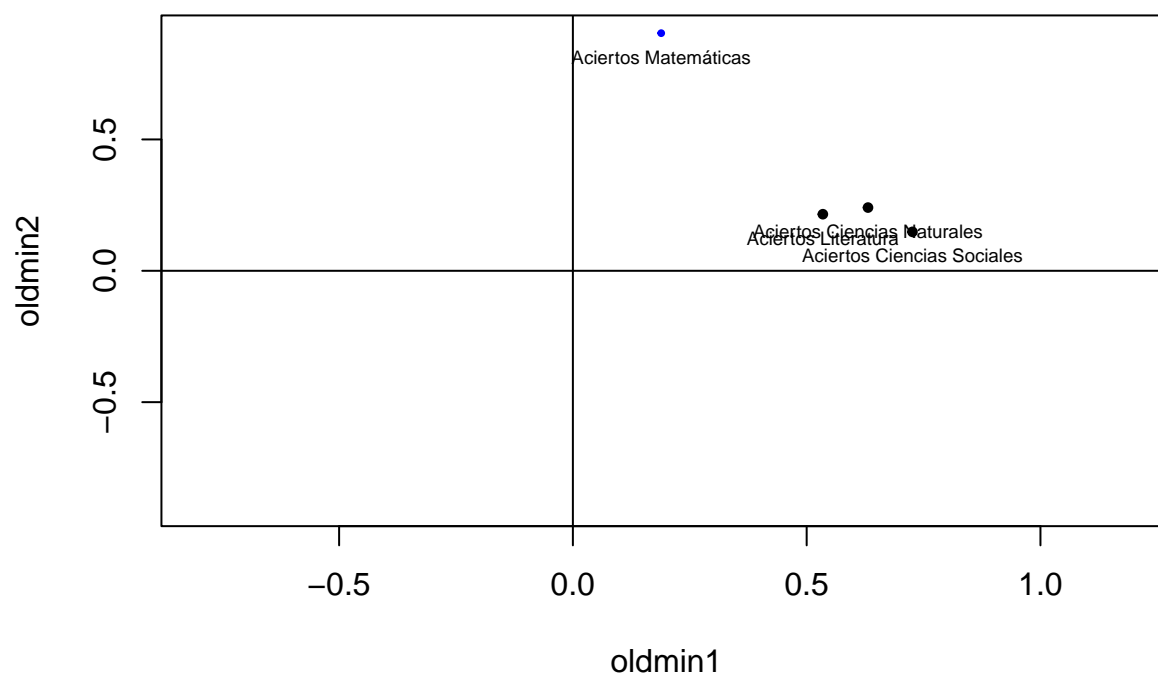
```
plot(fit.om.none, labels = row.names(R), cex=0.6, xlim=c(-0.8, 1.2), ylim=c(-0.9, 0.9),  
     main="Factor Analysis without Rotation")
```

Factor Analysis without Rotation

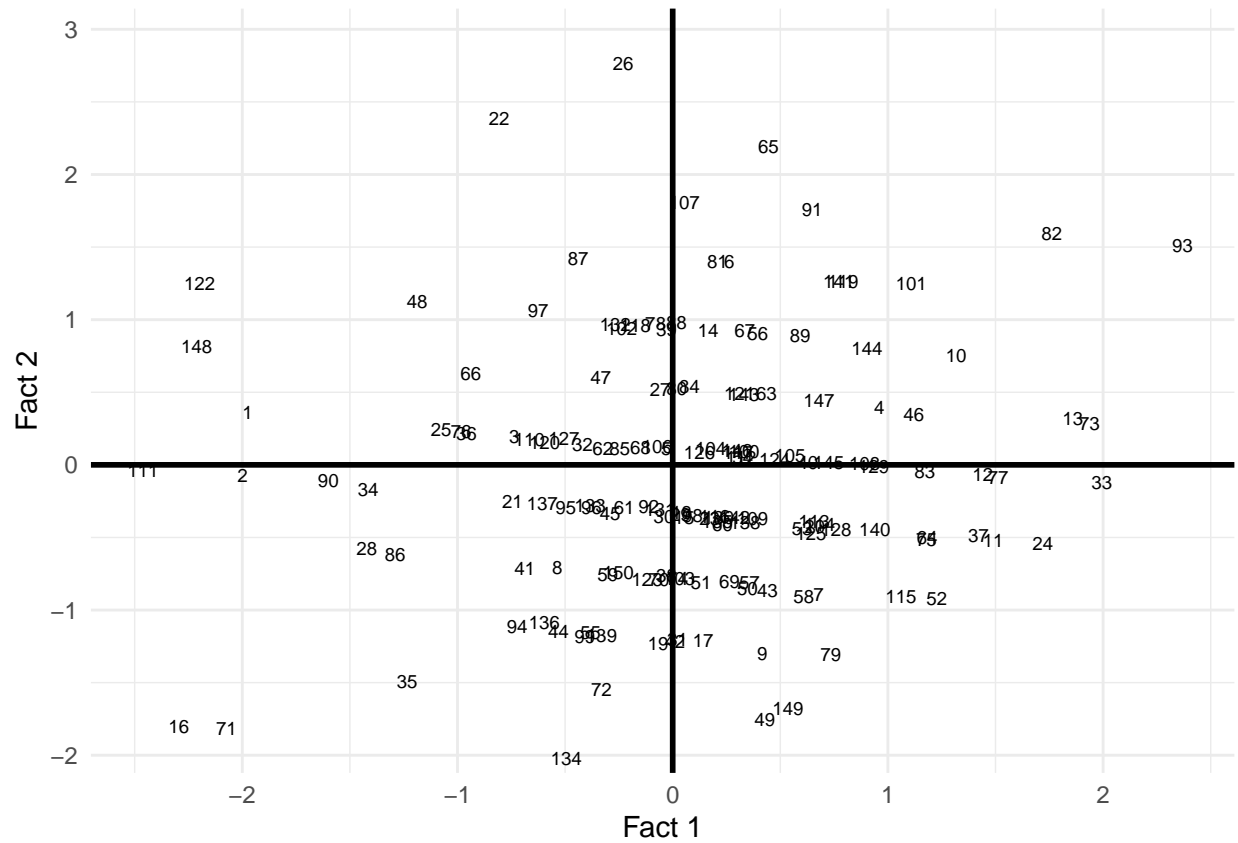


```
plot(fit.om.rot.pro, labels = row.names(R), cex=0.6, xlim=c(-0.8, 1.2), ylim=c(-0.9, 0.9),  
     main="Factor Analysis with Quartimax Rotation")
```

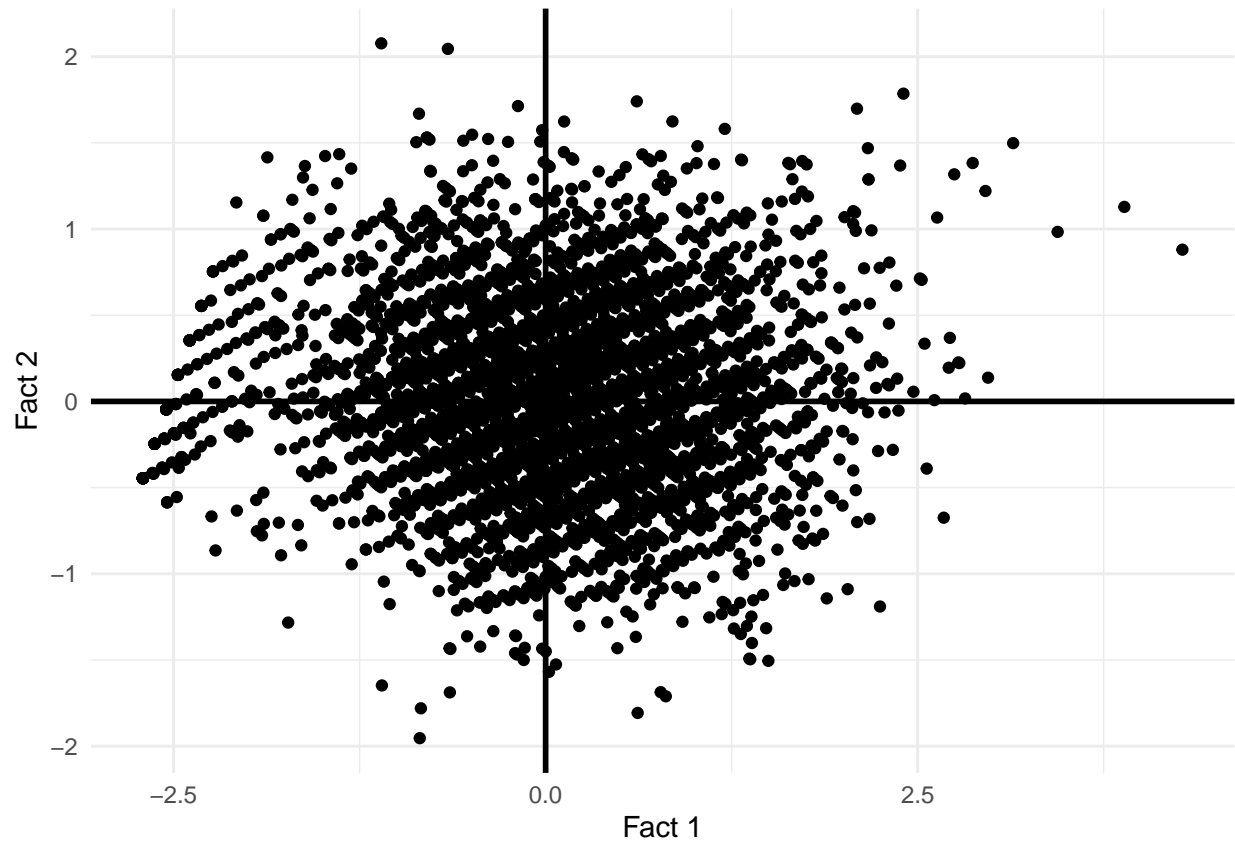
Factor Analysis with Quartimax Rotation



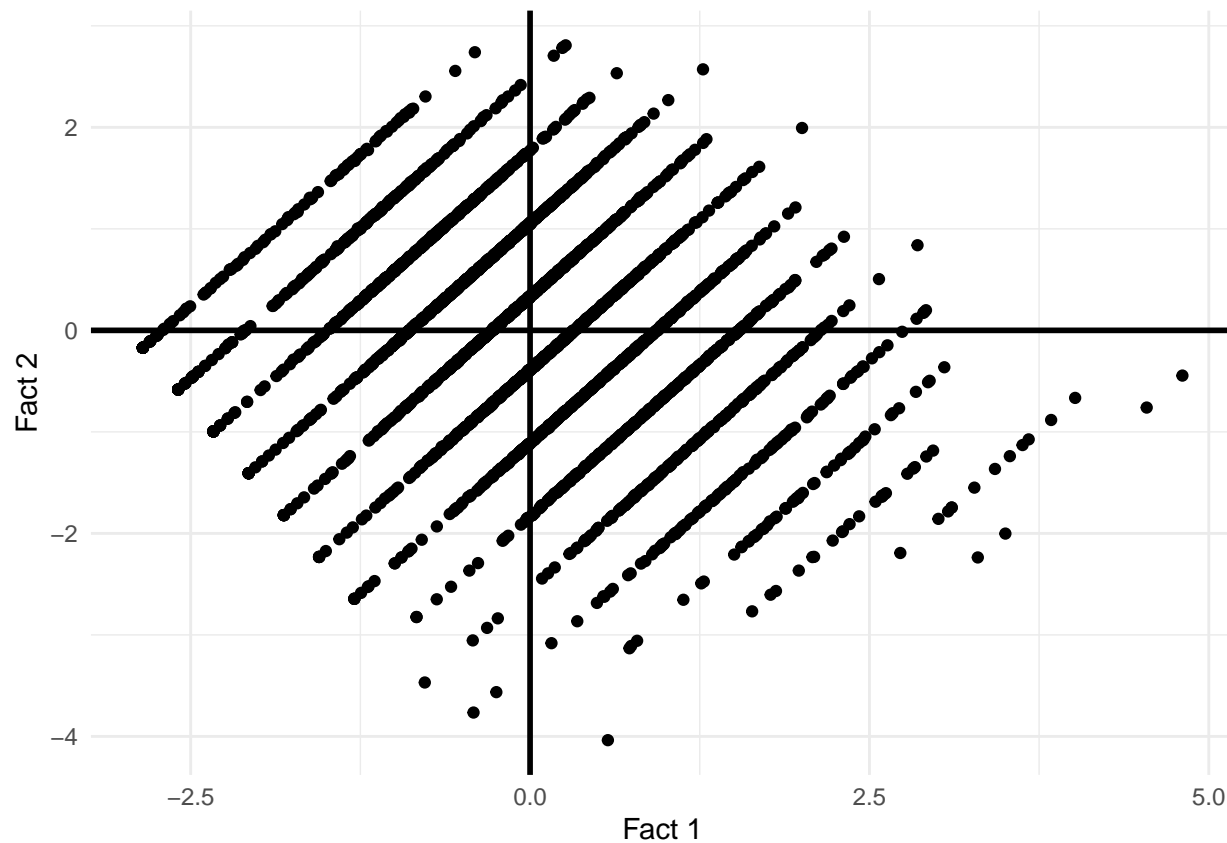
```
Fp=fit.om$scores
grap.fact=data.frame(y1=Fp[1:150,1],y2=Fp[1:150,2],lab=1:150,
                     grupo={dataTrab$Codigo.Estudiante}[1:150])
ggplot(grap.fact,aes(y1,y2,label=lab))+geom_text(vjust=2,size=2.35)+
  xlab("Fact 1")+ylab("Fact 2")+geom_hline(yintercept =0,size=1)+
  geom_vline(xintercept = 0,size=1)+theme_minimal()
```



```
Fp2=fit.ml$scores
grap.fact=data.frame(y1=Fp2[,1],y2=Fp2[,2],grupo=dataTrab$Codigo.Estudiante)
ggplot(grap.fact,aes(y1,y2))+geom_point()+
  xlab("Fact 1")+ylab("Fact 2")+geom_hline(yintercept =0,size=1)+
  geom_vline(xintercept = 0,size=1)+theme_minimal()
```



```
Fp3=fit.pa$scores
grap.fact=data.frame(y1=Fp3[,1],y2=Fp3[,2],grupo=dataTrab$Codigo.Estudiante)
ggplot(grap.fact,aes(y1,y2))+geom_point()+
  xlab("Fact 1")+ylab("Fact 2")+geom_hline(yintercept =0,size=1)+
  geom_vline(xintercept = 0,size=1)+theme_minimal()
```



Analisis de Conglomerados

Welcome! Want to learn more? See two factoextra-related books at <https://goo.gl/ve3WBa>

##

Attaching package: 'tidyr'

The following object is masked from 'package:RCurl':

##

complete

The following objects are masked from 'package:Matrix':

##

expand, pack, unpack

Warning: 'summarise_each()' was deprecated in dplyr 0.7.0.

Please use 'across()' instead.

Table 17: Matriz de tasas por cada provincia

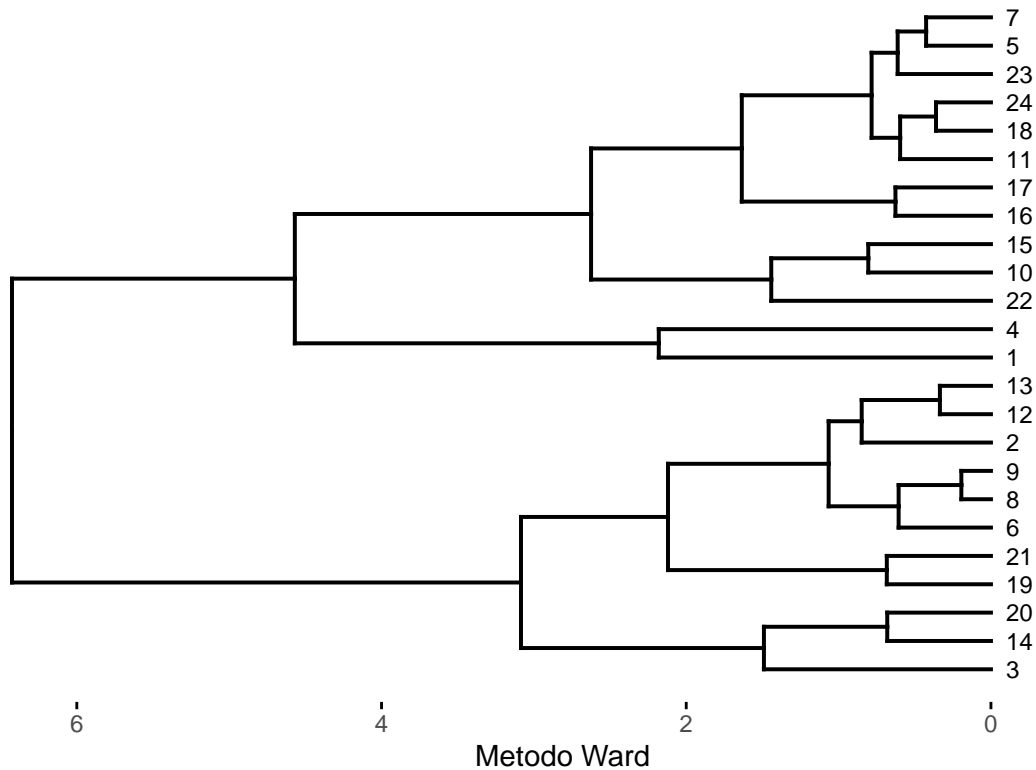
	Provincia	Aciertos Matemáticas	Aciertos Literatura	Aciertos Ciencias Naturales	Aciertos Ciencias Sociales
1	Azuay	4.33	9.33	12.33	8.00
2	Bolívar	4.40	6.80	10.50	7.14
3	Cañar	4.38	6.75	11.58	7.62
4	Carchi	6.11	9.11	13.56	8.22
5	Cotopaxi	4.72	7.69	11.61	7.39
6	Chimborazo	4.31	7.73	10.50	6.44
7	El Oro	4.58	7.82	11.24	7.33
8	Esmeraldas	4.23	7.27	10.40	6.65
9	Guayas	4.30	7.35	10.30	6.78
10	Imbabura	4.97	8.31	10.03	6.84
11	Loja	4.64	8.12	10.94	7.37
12	Los Ríos	4.33	6.85	10.33	6.57
13	Manabí	4.22	6.88	10.07	6.40
14	Morona Santiago	3.67	6.33	10.67	7.00
15	Napo	5.45	8.09	10.45	7.27
16	Pastaza	5.22	8.78	11.33	7.11
17	Pichincha	4.98	8.78	11.03	7.61
18	Tungurahua	5.02	7.91	11.15	7.27
19	Zamora Chinchipe	4.35	7.12	9.18	5.94
20	Galápagos	4.10	6.00	10.60	7.40
21	Sucumbíos	4.42	7.58	9.32	6.42
22	Orellana	4.33	8.00	9.50	7.50
23	Santo Domingo	4.89	8.20	11.58	7.37
24	Santa Elena	4.84	7.66	11.08	7.09

```
matriz_distancia<-dist(prov_set)
mat_distanciaH<-hclust(matriz_distancia,method = "ward.D2")
dendrogram<-as.dendrogram(mat_distanciaH)
graph_dendo<-fviz_dend(dendrogram,cex = 0.6,horiz = T,
                      main="Dendograma de las Provincias del Ecuador evaluadas",
                      ylab = "Metodo Ward")
```

```
## Warning: 'guides(<scale> = FALSE)' is deprecated. Please use 'guides(<scale> =
## "none")' instead.
```

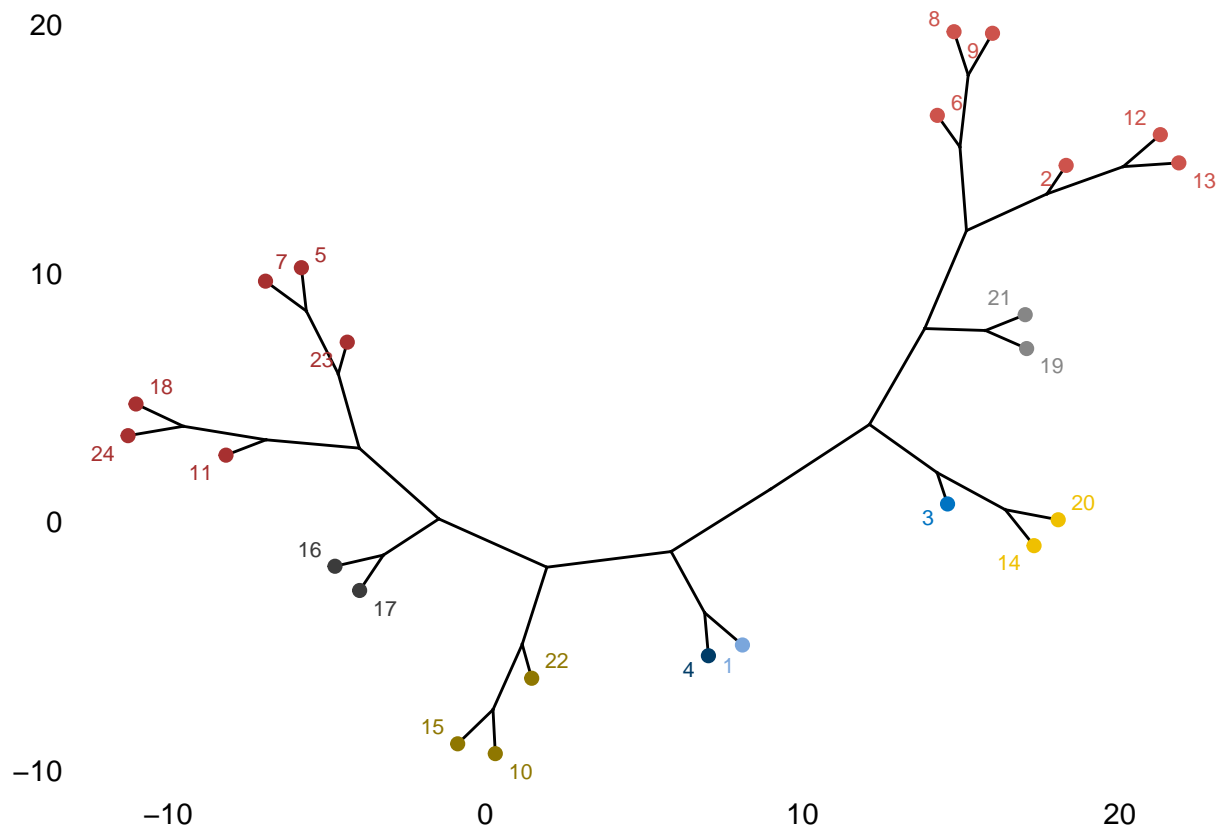
```
graph_dendo
```


Dendograma de las Provincias del Ecuador evaluadas



Tomando la cantidad de 9 clusters y mediante el metodo *Ward* obtenemos las siguientes agrupaciones:

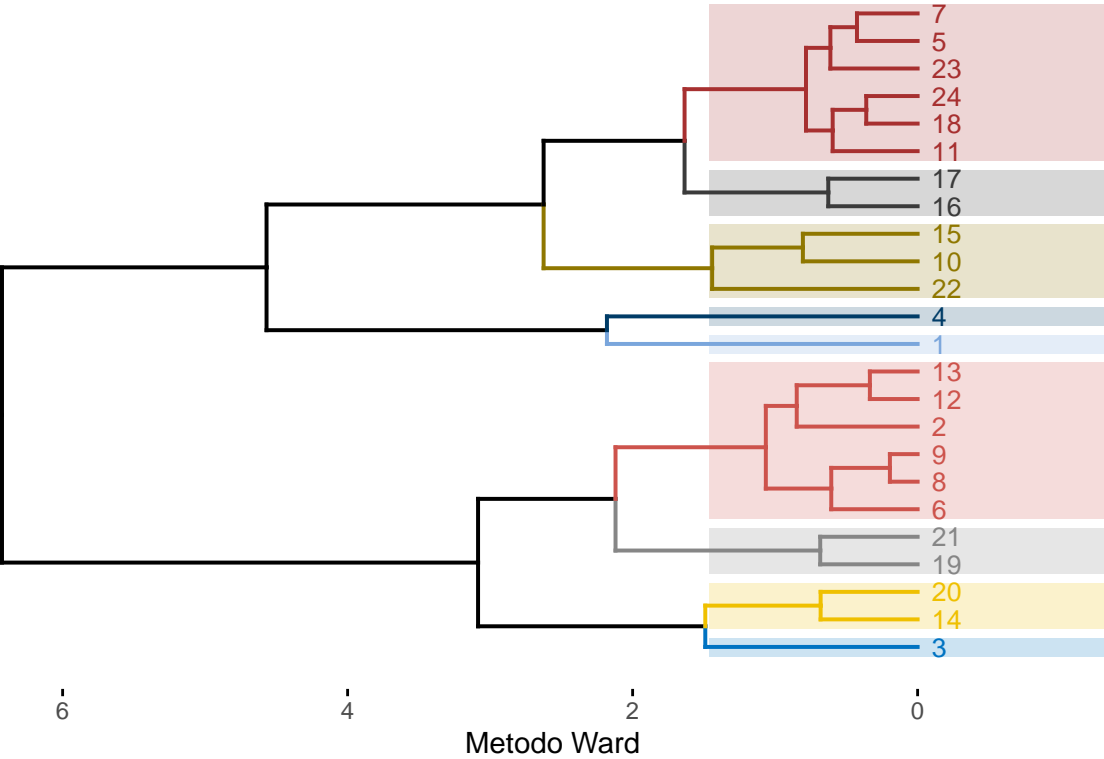
```
fviz_dend(dendogram,cex = 0.7,type = "phylogenetic",
          main="Dendograma de las Provincias del Ecuador evaluadas",
          ylab = "Metodo Ward",repel = T,k=9,k_colors = "jco")
```



```
fviz_dend(dendogram,cex = 0.7,horiz = T,rect = T,rect_border = "jco",rect_fill = T,
  main="Dendograma de las Provincias del Ecuador evaluadas",
  ylab = "Metodo Ward",repel = T,k=9,k_colors = "jco")
```

```
## Warning: 'guides(<scale> = FALSE)' is deprecated. Please use 'guides(<scale> =
## "none")' instead.
```

Dendograma de las Provincias del Ecuador evaluadas



CONCLUSIÓN:

Se puede concluir que analizar rendimientos académicos en los aspirantes que rinden el EADES abren una serie de interrogantes sobre la calidad en la educación secundaria del sistema educativo público ecuatoriano sobre todo en la asignatura de matemáticas y los items relacionados con la lógica. Los resultados motivan de forma obligada indagar más sobre los factores que causan el bajo rendimiento para superar las brechas y así mejorar los estándares de docencia y en especial el nivel de conocimientos adquiridos por los alumse puede concluir que analizar rendimientos académicos en los aspirantes que rinden el EAES abren una serie de interrogantes sobre la calidad en la educación secundaria del sistema educativo público ecuatoriano sobre todo en la asignatura de matemáticas y los items relacionados con la lógica.

BIBLIOGRAFÍA

1. Johnson, R. and Wichern, D., 2007. *Applied multivariate statistical analysis*. 6th ed. New Jersey: Pearson Education, Inc.
2. Batista Foguet, J. and Martínez Arias, M., 1989. *Análisis multivariante: análisis en componentes principales*. Barcelona: Hispano Europea.
3. Datar, R. and Garg, H., 2019. *Hands-On Exploratory Data Analysis with R*. Birmingham: Packt Publishing, Limited.
4. Romesburg, H., 2004. *Cluster analysis for researchers*. North Carolina: Lulu Press.
5. Aldenderfer, M. and Blashfield, R., 1984. *Cluster analysis*. Beverly Hills: Sage Publications.