Facultad de Ingeniería
UNER

Tecnicatura Universitaria en Procesamiento y Explotación de Datos.

Análisis de desempeño de estudiantes en el Departamento de Colón, E.R. durante el año 2022.

Por

Armú Yamil, Zamora Rodrigo

14/06/2023

Desarrollado como resultado del análisis llevado a cabo sobre las escuelas del departamento de Colón en la asignatura Exploración de Datos Multivariados.

El presente trabajo es el resultado del análisis llevado a cabo sobre el rendimiento académico de los estudiantes en las escuelas del departamento de Colón.

Con los datos disponibles se realizaron técnicas de limpieza y organización, estudios de representatividad de la muestra, estudios sobre las notas obtenidas por los alumnos de diferentes años y materias, comparaciones entre los mismos resultados, y análisis de correlación de los datos. Todo mediante técnicas estadísticas que permitieron el análisis, acompañado de la elaboración de representaciones gráficas.

Con el fin de diagnosticar el estado general de la educación común y técnica, además de recabar información de utilidad y analizar en particular el rendimiento de los estudiantes en materias básicas de secundaria y primaria, cursadas en el año 2022. Para luego así asesorar en la implementación de programas complementarios.

Los datos proporcionados para llevar a cabo el presente trabajo fueron recolectados de segunda mano, pertenecientes a los boletines digitales de los estudiantes del departamento de Colón.

El análisis pertinente de los datos, se llevó a cabo mediante las siguientes herramientas informáticas:

- Lenguaje de programación R y diversas librerías
- Interfaz Jupyter Notebook v6.5.2
- Visual Studio Code
- Google Workspace (Meet, Drive, Docs, etc)

Acondicionamiento de los datos.

La base de datos asignada sobre el departamento de Colón, presentaba en primer lugar nombres de variables en diferentes notaciones y con errores ortográficos, se hallaron además datos mal cargados, con formatos incorrectos, campos vacíos y duplicados.

En la etapa de acondicionamiento del dataset para el análisis solicitado, se corrigieron los errores mencionados anteriormente, además se filtraron las variables significativas y los datos relevantes, así también se realizó una homogenización de las asignaturas iguales pero con diferente nombre, igualmente se unieron los registros de notas correspondientes a diferentes trimestres de una asignatura en particular como un solo registro, de la siguiente forma: (“NotaPrimerT”, “NotaSegundoT”, “NotaTercerT”), como también promedio (“Promedio”) y condición final (“Resultado”). Vale mencionar que se mantuvieron solo los registros completos, es decir, que contengan notas cargadas en los tres trimestres.

Estudio de representatividad de los datos proporcionados con respecto al departamento.

Primaria

Matrícula total en departamento	Matrícula cargada en notas	Porcentaje de representatividad	Mínimo muestral
8610	4596	53.38	368

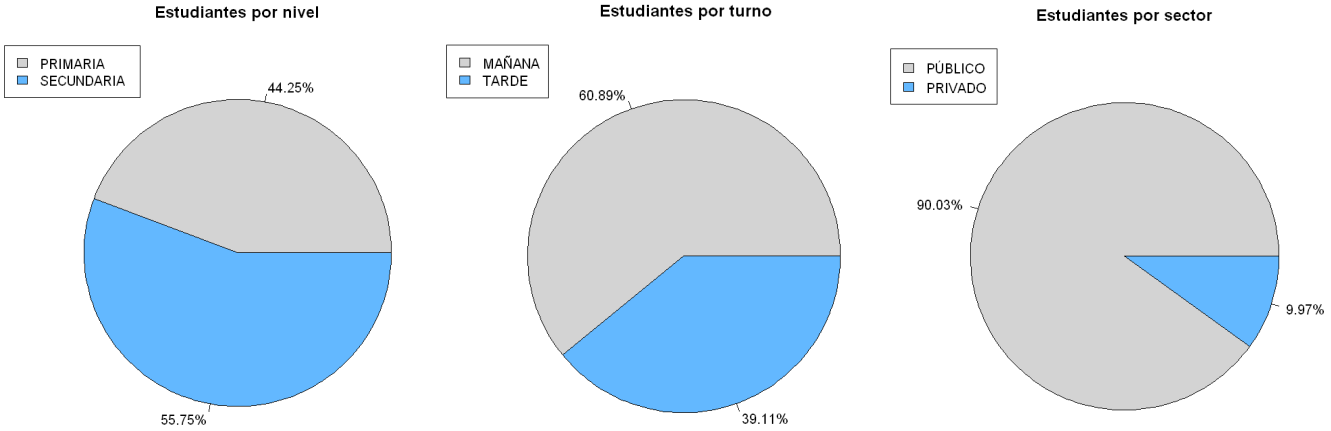
Secundaria

Año	Matrícula total en departamento	Matrícula cargada en notas	Porcentaje de representatividad	Mínimo muestral
1	1541	1262	81.89	308
2	1508	1244	82.49	307
3	1273	955	75.02	296
4	1422	881	61.95	303
5	1095	697	63.65	285
6	1005	653	64.98	279
7	112	99	88.39	87

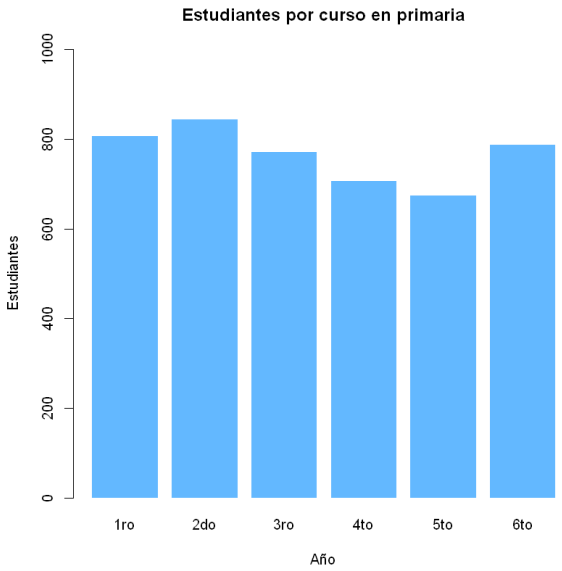
Análisis sobre la distribución de estudiantes en el departamento.

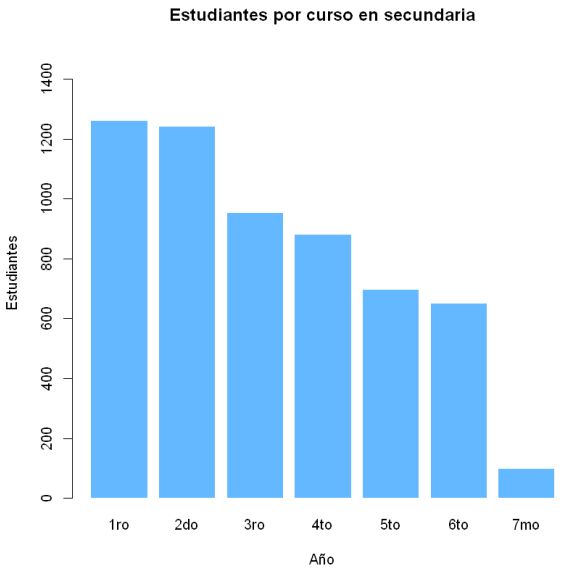
Se realizó un análisis sobre la distribución de estudiantes según nivel (primaria/secundaria), turno (mañana/tarde) y sector (público/privado).

Donde se observó un mayor número de estudiantes en secundaria (57%), prevalencia en el turno mañana (60%) y sector público (90%) debido a que no se registraron datos de primaria perteneciente al sector privado.



Se crearon gráficos para visualizar la distribución de estudiantes por año, tanto para primaria como secundaria.





Para el conjunto de primaria observamos una distribución homogénea, siendo segundo año el de mayor número de estudiantes.

Por el contrario para el conjunto de secundaria logramos observar una distribución concentrada en los primeros dos años, a los cuales les corresponde el máximo de alumnos, y menor para los siguientes.

Medidas de centralidad y dispersión correspondientes a los promedios de estudiantes de secundaria en materias como Matemática y Lengua.

Podemos observar que la media en ambas materias es similar en todos los años escolares, aunque la media en Lengua y Literatura tiende a ser ligeramente más alta que en Matemáticas en algunos años escolares. Por ejemplo, en el quinto año escolar, la media en Lengua y Literatura fue de 7.24, mientras que en Matemáticas fue de 6.80.

La mediana nos permite saber que más del 50% de los estudiantes de los diferentes años tuvieron promedio mayor a 6.

Matemática

año	media	mediana	moda	desvío estándar
1	6.54	6.33	6.00	1.63
2	6.24	6.00	6.00	2.06
3	6.31	6.33	6.00	1.86
4	6.58	6.33	6.00	2.04
5	6.80	7.00	6.00	1.72
6	7.26	7.33	6.00	1.81

Lengua y Literatura

año	media	mediana	moda	desvío estándar
1	6.80	6.67	6.00	1.40
2	6.47	6.33	6.00	1.64
3	6.66	6.67	6.00	1.71
4	6.41	6.33	6.00	1.68
5	7.24	7.33	6.00	1.51

En términos de la desviación estándar, podemos observar que la variabilidad en Matemáticas es generalmente mayor que en Lengua y Literatura. En particular, el desvío estándar en Matemáticas fue mayor en el segundo y cuarto año escolar, mientras que en Lengua y Literatura el desvío estándar fue mayor en el tercer año escolar. En resumen, aunque la media en ambas materias es similar, la variabilidad en los resultados es mayor en Matemáticas en general.

Comparación de las notas de secundaria del primer año de las materias Matemática y Lengua y Literatura.

los valores de la media mediana y moda se corresponden ya que son similares.

la media, es representativa para ambas ya que las variables casi no tienen valores extremos y la media comparte valores similares con la mediana y moda.

la dispersión de los datos al ver los box plot tienen una dispersión normal ya que sus cajas son cuadradas un poco más achatadas, lengua tiene menos dispersión que matemática. La desviación estándar en matemática 1.6 y lengua 1.39.

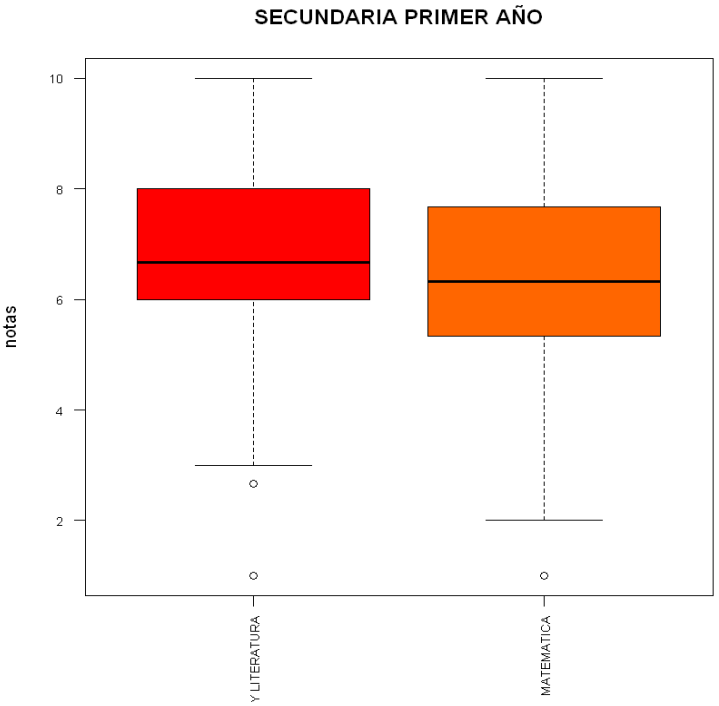
	Matemática	Lengua y Literatura
Muestra	1231	1179
Media	6.5	6.8
Mediana	6.3	6.67
Moda	6	6
Normalidad	No	No
Desvío estándar	1.6	1.39
IQR	2.3	2
Q3	7.6	8

Conclusión de las materias de lengua y matemática en primer año secundaria

las materias tienen una dispersión normal, matemática al tener mayor desvío estándar y un mayor rango intercuartil (matemática: 2.3 lengua: 2) tiene una mayor variabilidad.

Existen dos outliers en lengua y literatura y uno de matemática con notas muy bajas menores a 3 en una muestra de 1179 (lengua) y 1231 (matemática) lo cual es poco para una muestra tan grande.

considerando que el primer cuartil en la materia lengua da 6, por ende como mínimo el 75% de los alumnos aprobaron la materia, si considero que el rendimiento fue óptimo esa materia. En matemática no sucede lo mismo ya que para empezar el bigote inferior llega a valores más bajos que el de lengua, también la mediana es menor, de igual forma la mediana sigue estando por arriba de 6 por lo que se puede saber existen más aprobados que desaprobados.

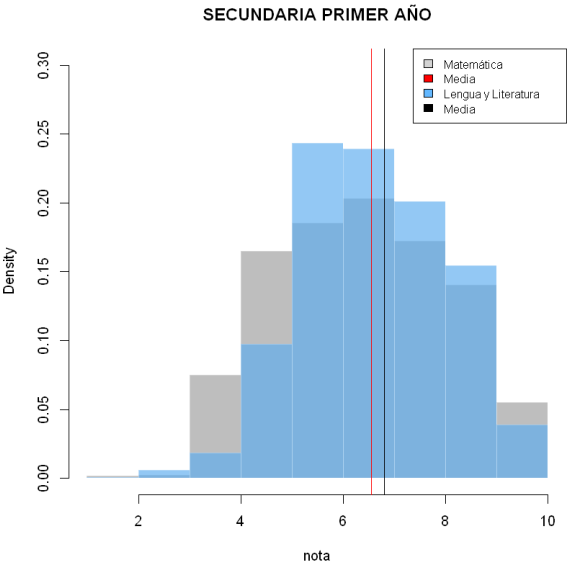


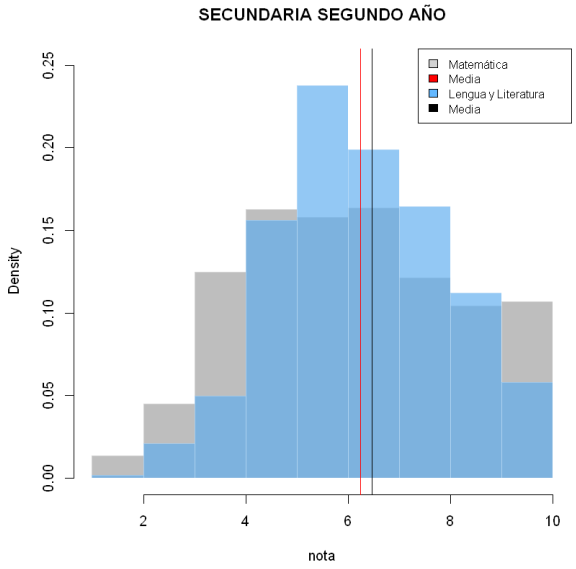
En conclusión lengua tuvo un promedio de notas óptimo en sus alumno y matemática no tanto, sin embargo no es malo el rendimiento ya que los desaprobados, sus notas se concentran en notas de 5 y 4 y solo hay 1 outlier por debajo de 3.

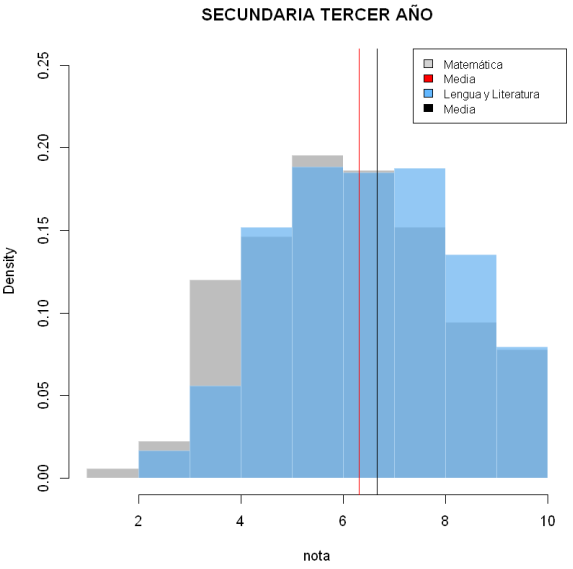
Se realizaron histogramas con los promedios de materias como matemática y lengua y literatura de secundaria, para cada uno de los diferentes años.

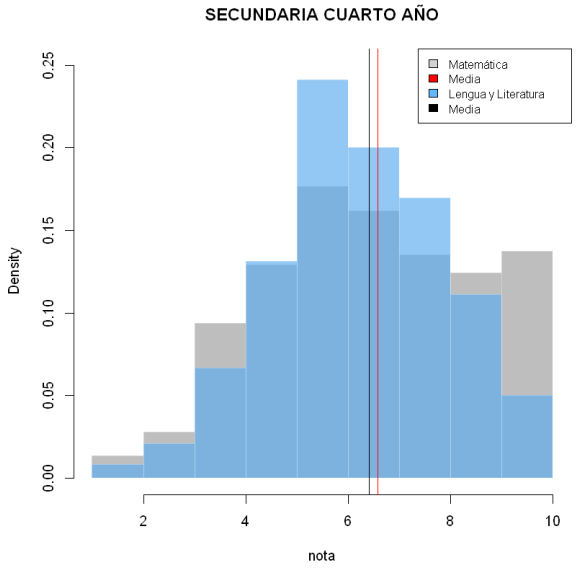
Se analizó la distribución, dispersión y morfología de estos. En general para todos los años la media resultó ser representativa del conjunto, además se visualizó alta frecuencia de promedios en torno al intervalo que contiene 6 considerado aceptable, pero también frecuencias altas para intervalos que contienen promedios inferiores a los mencionados.

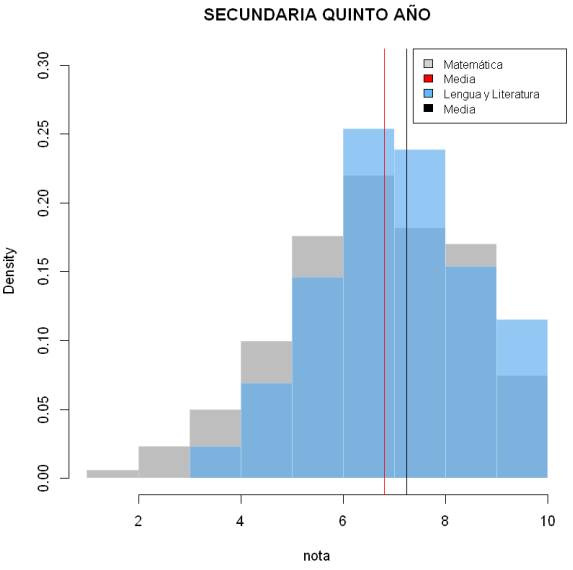
El mismo análisis se llevó a cabo para estas asignaturas pertenecientes a la educación primaria. En general se observó un desempeño superior con respecto a los promedios, donde por el contrario casi no se registraron casos donde estos fueran inferiores a 6.











Análisis de correlación lineal entre matemática y lengua y literatura

Se realizaron test de normalidad con lillie.test de las materias en los distintos años y todos dieron con valores muy bajos, números mucho menores a 0.05 entonces se decide usar test no paramétricos como spearman.

Secundaria

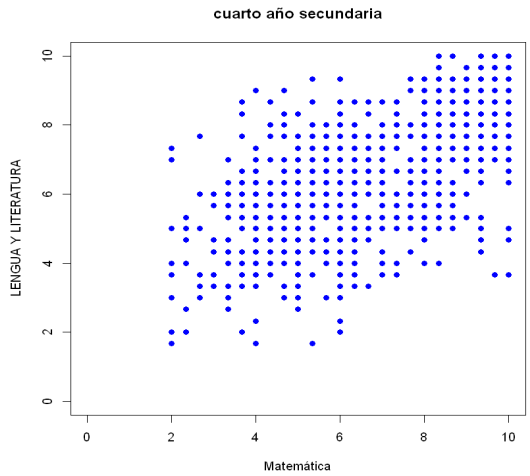
	primero	segundo	tercero	cuarto	quinto
rho	0.7696085	0.6664015	0.6594309	0.5775122	0.5245301
correlación	buena	buena	buena	moderada	moderada
p-valor	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16

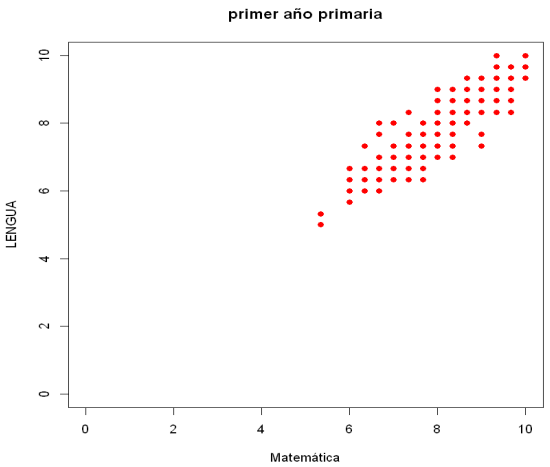
Utilizamos el método de correlación de Spearman para realizar una prueba de hipótesis para evaluar si hay una asociación monotónica significativa (positiva o negativa) entre las dos variables.

Primaria

	primero	segundo	tercero	cuarto	quinto	sexto
rho	0.9497736	0.92351	0.8929307	0.9145661	0.9338398	0.8872828
correlación	buena	buena	buena	buena	buena	buena
p-valor	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16

El primer gráfico es donde se obtuvo la mayor dispersión y el segundo la menor de las relaciones trabajadas.





conclusión: la relación entre las materias en primaria no varía y es muy buena, Mientras que en la secundaria más avanzado sea el año menor es la relación lineal algo similar sucede en primaria pero no es tan notorio. Como rho es cercano a 1, podemos interpretarlo como una correlación positiva fuerte entre matemática y lengua. El valor del p-valor en todos los años fue por debajo de 0,05. Esto indica que hay una fuerte evidencia en contra de la hipótesis nula de que no hay correlación entre matemática y lengua.

Análisis de los porcentajes de aprobados en las materias de cada año de la secundaria

Podemos observar que el porcentaje de aprobados en las diferentes materias varía significativamente en los resultados que se presentan. En general, podemos decir que las materias con los porcentajes de aprobación más altos son Educación Física, Geografía y Química, mientras que Matemáticas e Inglés tienen los porcentajes más bajos.

Es interesante notar que la materia de Biología tiene resultados bastante dispares, con porcentajes que van desde el 66.3% hasta el 92.2%. La materia de Lengua y Literatura también presenta una variabilidad significativa en sus resultados.

Asignatura /Año	Primero	Segundo	Tercero	Cuarto	Quinto	Sexto
Biología	74.9%	68.9%	66.3%	71.3%	79.1%	92.1%
Educación Física	95.7%	92.9%	94.4%	95.5%	97.3%	97.7%
Fisicoquímica	79.8%	72.1%	75.1%			
Geografía	84.1%	71.4%	70.4%	73.6%	86.7%	93.9%
Historia	79.9%	74.5%	72.2%	82.8%		
Inglés	69.8%	77.3%	73.3%	66.6%	79.8%	
Lengua y Literatura	77.5%	64.2%	68.0%	65.1%	83.0%	
Matemática	66.5%	58.1%	61.2%	63.3%	75.5%	80.4%
Física				72.5%	80.4%	80.5%
Química				69.7%	80.5%	89.6%

Análisis de los alumnos con malas notas que pudieron aprobar la materia.

MATEMÁTICA	3.92%
LENGUA Y LITERATURA	3.99%
EDUCACIÓN FÍSICA	20.95%
HISTORIA	2.61%
BIOLOGÍA	2.52%
GEOGRAFÍA	4.15%
INGLÉS	5.92%

La tabla contiene el porcentaje de alumnos de secundaria que pudieron aprobar una materia habiendo desaprobado el primer y segundo trimestre de la misma.

Podemos observar que Educación Física presenta un porcentaje considerablemente superior al resto, lo cual nos indica que a los estudiantes que desaprobaron los primeros dos trimestres les resultó más fácil aprobar esta asignatura a comparación de las otras.

Hipótesis de correlación negativa entre matemática y educación física.
Con el fin de responder a la hipótesis planteada, la cual afirma que los estudiantes que obtienen notas superiores en matemática, por el contrario obtienen notas inferiores en educación física, se realizó un análisis de correlación de los promedios de los estudiantes entre ambas asignaturas.

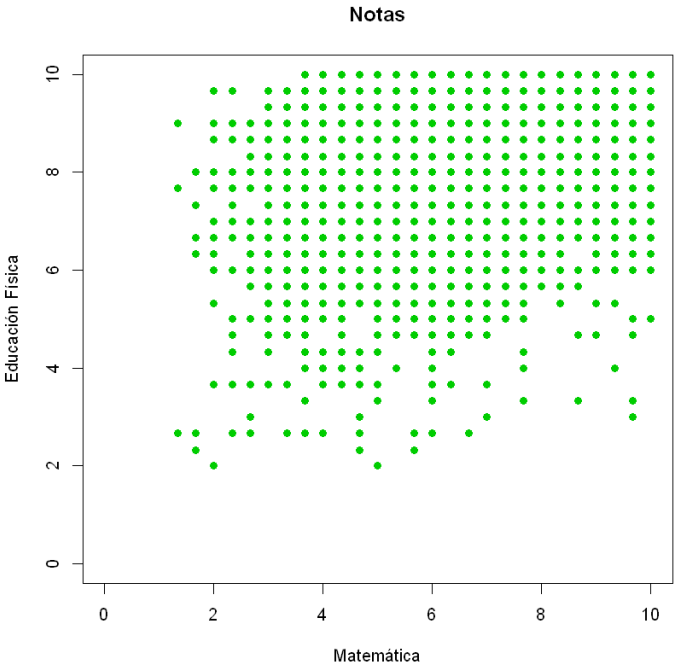
En primer lugar se comprobó que ambas variables tengan una cantidad similar de datos, y se crearon histogramas que permitieron visualizar las respectivas distribuciones.
Se realizaron pruebas de normalidad, que, debido a la cantidad de datos para cada variable, de 5155 y 5369, para matemática y ed. física respectivamente, se utilizó el método Lilliefors (Kolmogorov-Smirnov).

variable	D	p-value
matemática	0.054355	< 2.2e-16
ed. física	0.1213	< 2.2e-16

H_0 : La variable sigue una distribución normal.

Con un p-value inferior a 0,05 rechazamos la hipótesis nula (H_0).

Se procede a realizar un estudio de correlación con el método de Spearman, y los resultados obtenidos arrojan una correlación positiva débil, con un valor rho de 0.3343801 y p-value < 2.2e-16.



Se repite el análisis, ésta vez se aplicaron otros filtros sobre los datos (turno mañana, modalidad común, escuela pública), a fin de eliminar variables que podrían estar interfiriendo en los resultados esperados.

variable	D	p-value
matemática	0.11484	3.09e-16
ed. física	0.12256	< 2.2e-16

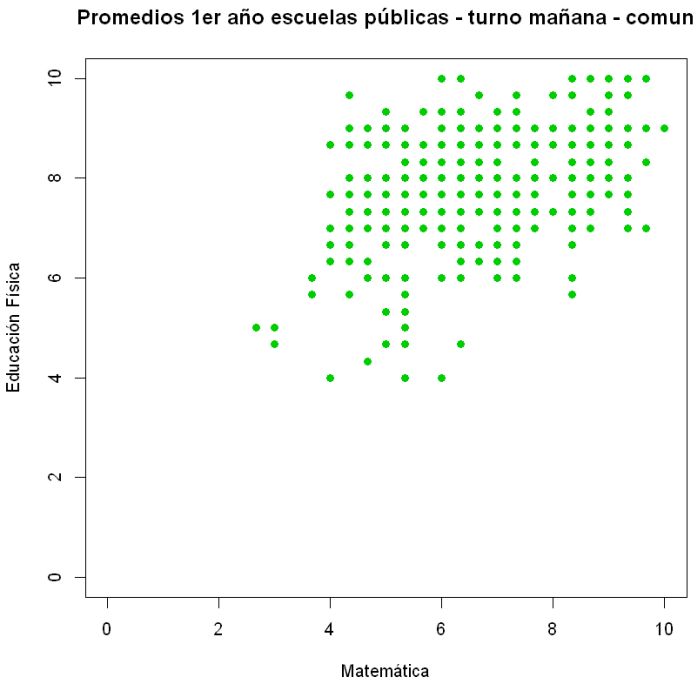
H_0 : La variable sigue una distribución normal.

Con un p-value inferior a 0,05 rechazamos la hipótesis nula (H_0).

Se aplica el test de correlación de Spearman en la segunda muestra, con un tamaño inferior.

Los resultados arrojan que la correlación ha aumentado con un valor rho = 0.427 y p-value < 2.2e-16, con lo cual se evidencia la inexistencia de correlación negativa, y se rechaza la hipótesis planteada inicialmente.

Se logra observar el cambio en el comportamiento de la nueva gráfica.



Análisis de notas mediante Comparación de Medias y Tipificación de Valores.

En éste apartado realizamos un análisis sobre los datos referidos a notas en asignaturas como Matemática y Lengua en estudiantes de nivel secundario. Se realizaron diferentes test (paramétricos y no paramétricos) con el objetivo de probar las distintas hipótesis planteadas.

Si un estudiante de tercer año de una escuela de su departamento obtuvo un 5 en Matemática, ¿cómo le fue respecto a los puntajes obtenidos en su curso?

Para responder a la pregunta planteada, en primer lugar filtramos los datos de una escuela en particular, referidos a la asignatura Matemática y correspondientes a tercer año.

Con estos datos calculamos el valor de la media aritmética y el desvío estándar para la variable "Promedio", luego utilizando tipificación calculamos el valor **z** para la nota del estudiante en cuestión mediante la fórmula:

$$Z = \frac{X - \mu}{\sigma}$$

y obtuvimos un valor z = -2.75. A partir de esto pudimos contrastar la nota obtenida por el alumno en particular con las notas obtenidas por todos los alumnos de su escuela, permitiéndonos observar que su rendimiento en la asignatura Matemática está 2.75 desvíos estándares por debajo del promedio del curso.

¿Podría afirmar que la media de su departamento en las notas de Matemática de primer año estuvo por encima de 7?

En primer lugar identificamos la variable aleatoria de interés para esta prueba, y planteamos la hipótesis nula (H_0) y la hipótesis alternativa (H_a) correspondientes.

X: Los promedios de matemática de los estudiantes de 1er año de secundaria del Departamento de Colón.

H_0: En promedio las notas de los estudiantes son mayores o iguales a 7;

H_a: En promedio las notas de los estudiantes son menores a 7.

Luego verificamos si los datos involucrados en la prueba cumplen los supuestos de normalidad, debido a la cantidad utilizamos una prueba de normalidad de Lilliefors (Kolmogorov-Smirnov) la cual arrojó los valores D = 0.071411, p-value = 3.402e-16. Permitiéndonos observar que aparentemente los datos no siguen una distribución normal.

En consecuencia a los resultados obtenidos, procedemos a probar la hipótesis planteada mediante una prueba unilateral izquierda (cola izquierda) no paramétrica de Wilcoxon, con un nivel de confianza de 95%, obteniéndose así los valores V = 231460, p-value < 2.2e-16. A partir de los resultados, se rechaza la hipótesis nula (H_0) a favor de la hipótesis alternativa (H_a).

¿El 50% de los estudiantes de su departamento obtiene en Lengua notas inferiores a 6?

En primer lugar, para responder a la pregunta seleccionamos el 50% central de los datos correspondientes a la asignatura Lengua. Luego al igual que en pruebas anteriores, se identifican la variable aleatoria de interés para esta prueba, la hipótesis nula (H_0) y la hipótesis alternativa (H_a) correspondientes.

X: Los promedios de lengua del 50% de los estudiantes de secundaria del Departamento de Colón.

H_0: En promedio las notas de los estudiantes son menores o iguales a 6;

H_a: En promedio las notas de los estudiantes son mayores a 6.

Realizamos una prueba de normalidad de Lilliefors (Kolmogorov-Smirnov) la cuál arrojó los valores D = 0.14775, p-value < 2.2e-16, la cuál nos informa que los datos no cumplen los supuestos de normalidad.

Procedemos a realizar una prueba unilateral derecha no paramétrica de Wilcoxon, con un nivel de confianza de 95%, el valor del estadístico de la prueba es V = 1483165 con un p-value < 2.2e-16.

A partir de los resultados, se rechaza la hipótesis nula (H_0) a favor de la hipótesis alternativa (H_a).

Es cierto que, en su departamento, en todos los años del sistema educativo, ¿las notas que se obtienen en el tercer trimestre son superiores a las del primer trimestre?

Identificamos la variable aleatoria de interés para esta prueba, la hipótesis nula (H_0) y la hipótesis alternativa (H_a) correspondientes.

X: Las notas del primer trimestre y tercer trimestre de los estudiantes de secundaria del Departamento de Colón.

H_0: Las notas del tercer trimestre son mayores a las del primero;

H_a: Las notas del tercer trimestre no son mayores a las del primero.

Realizamos una prueba de normalidad de Lilliefors (Kolmogorov-Smirnov) la cuál arrojó los valores D = 0.17307, p-value < 2.2e-16, la cuál nos informa que los datos no cumplen los supuestos de normalidad.

Procedemos a realizar una prueba unilateral izquierda no paramétrica de Wilcoxon, con un nivel de confianza de 95%, el valor del estadístico de la prueba es V = 137851975, con un p-value = 1.

A partir de los resultados, no se rechaza la hipótesis nula (H_0).

¿Hay diferencias significativas entre el rendimiento en Matemática de primer año del turno tarde y del turno mañana?

Identificamos la variable aleatoria de interés para esta prueba, la hipótesis nula (H_0) y la hipótesis alternativa (H_a) correspondientes.

X: Las notas de Matemática del turno tarde y turno mañana de los estudiantes de primero de secundaria del Departamento de Colón.

H_0: las notas del turno mañana no difieren de las del turno tarde;

H_a: las notas del turno mañana difieren de las del turno tarde.

Realizamos pruebas de normalidad de Lilliefors (Kolmogorov-Smirnov) para los datos del turno mañana y tarde, las cuáles arrojaron valores D = 0.071847, p-value = 1.085e-08, y D = 0.083244, p-value = 4.67e-07 respectivamente, lo cuál nos informa que los datos no cumplen los supuestos de normalidad.

Procedemos a realizar una prueba bilateral (dos colas) no paramétrica de Wilcoxon, con un nivel de confianza de 95%, el valor del estadístico de la prueba es W = 139684, con un p-value = 0.2612.

A partir de los resultados, no se rechaza la hipótesis nula (H_0).

En conclusión, mediante el análisis realizado en ésta sección y tras responder a las preguntas planteadas, podemos recopilar información sobre las notas de los estudiantes de nuestro departamento. Se destaca: La media de las notas de Matemática de primer año estuvo por debajo de 7; El 50% de los estudiantes en el departamento obtiene en Lengua notas superiores a 6; En el departamento para todos los años, las notas que se obtienen en el tercer trimestre son superiores a las del primer trimestre; Por último no existen diferencias significativas entre el rendimiento en Matemática de primer año del turno tarde respecto del turno mañana.

Análisis de la matrícula de estudiantes en nivel secundario.

Para este apartado se analizó la matrícula de estudiantes por ciclo en el nivel secundario, se distinguieron dos grupos, el ciclo básico: conformado por los datos de matrícula correspondientes a 1^{er}, 2^{do} y 3^{er} año, y por otro lado el ciclo orientado: conformado por 4^{to}, 5^{to} y 6^{to} año.

El fin es contrastar la matrícula de ambos ciclos para probar si existen diferencias significativas entre sí, esto mediante distintos test de pruebas de hipótesis y representaciones gráficas de los datos como boxplots.

En primer lugar, a partir del dataset “Secundaria3”, analizamos si la matrícula en promedio del ciclo básico difiere significativamente de la del ciclo orientado.

Mediante tests de normalidad de Shapiro-Wilk, los cuales arrojaron valores p = 01395 y p = 0.002684, para el ciclo básico y ciclo orientado respectivamente, comprobamos que los grupos no siguen una distribución normal.

Tras ello, se procedió a realizar un test no paramétrico de Wilcoxon de dos colas, con un nivel de confianza de 95%, para comprobar la hipótesis establecida.

H_0: Las matrículas entre ciclo básico y ciclo orientado no difieren significativamente.

H_a: Las matrículas entre ciclo básico y ciclo orientado si difieren significativamente.

Los resultados de la prueba arrojaron un valor estadístico W = 502, y un p-value = 0.2078, ya que este último es superior a 0.05 no es posible rechazar la hipótesis nula, es decir las matrículas no difieren significativamente entre ambos grupos.

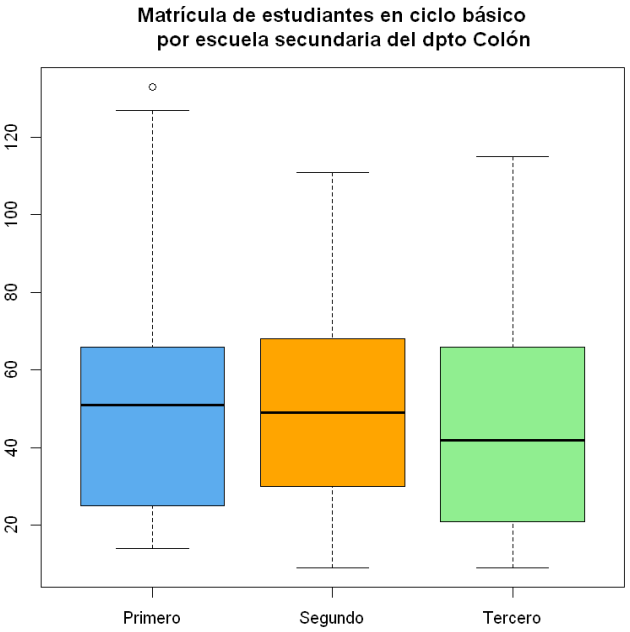
Luego, haciendo uso del mismo dataset se analizó cada grupo por separado, buscando diferencias significativas entre los cursos que componen cada uno.

COMPARACIÓN DE MEDIAS

Análisis de Ciclo Básico.

En primer lugar se obtienen descriptivos sobre los datos de matrícula del ciclo básico, y se creó un boxplot de los mismos.

Primero	Segundo	Tercero
Min. : 14.00	Min. : 9	Min. : 9.0
1st Qu.: 25.00	1st Qu.: 30	1st Qu.: 21.0
Median : 51.00	Median : 49	Median : 42.0
Mean : 53.14	Mean : 52	Mean : 43.9
3rd Qu.: 66.00	3rd Qu.: 68	3rd Qu.: 66.0
Max. :133.00	Max. :111	Max. :115.0



Se observan diferencias en la media aritmética y la mediana de Tercero con respecto a Primero y Segundo. Además una variabilidad similar entre los tres años, observable en el tamaño de las cajas en el gráfico.

Los tests de normalidad de Shapiro-Wilk correspondientes revelaron que los datos no cumplen con los supuestos de normalidad, por lo tanto procedimos a realizar un test no paramétrico U de Kruskal-Wallis con un nivel de confianza del 95% para contrastar si las diferentes muestras correspondientes a cada año están equidistribuidas.

H_0: Todas las muestras provienen de la misma distribución.

H_a: Al menos una de las muestras proviene de una distribución diferente.

La prueba realizada arrojó los siguientes resultados; Kruskal-Wallis chi-squared = 1.2723, df = 2, p-value = 0.5293. Con un valor p > 0.05 no se rechaza la hipótesis nula (H_0: Todas las muestras provienen de la misma distribución), es decir que mediante la prueba realizada no se obtienen diferencias significativas.

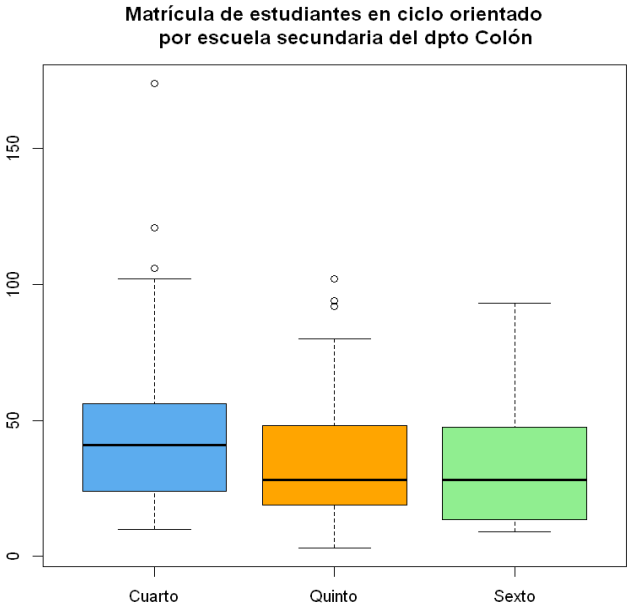
Se calcula la potencia del test ANOVA con el fin de estimar la potencia del test U de Kruskal-Wallis realizado obteniéndose una potencia de 0.524, Se estima que el test realizado (Kruskal-Wallis) posee una potencia menor. Esto sugiere que quizás si existan diferencias significativas, solo que nuestra prueba no es lo suficientemente potente para detectarlas.

Análisis de Ciclo Orientado.

Pasando al análisis del ciclo orientado, se busca comprobar la existencia de diferencias significativas en la matrícula de los diferentes años que conforman el grupo (4^{to}, 5^{to}, 6^{to}), para ello primero obtuvimos descriptivos del conjunto de datos y elaboramos un boxplot a fin de representarlos.

Cuarto	Quinto	Sexto
Min. : 10.00	Min. : 3.00	Min. : 9.00
1st Qu.: 24.00	1st Qu.: 19.00	1st Qu.:13.75
Median : 41.00	Median : 28.00	Median :28.00
Mean : 49.03	Mean : 37.76	Mean :35.89
3rd Qu.: 56.00	3rd Qu.: 48.00	3rd Qu.:47.25
Max. :174.00	Max. :102.00	Max. :93.00
		NA's :1

Se observan diferencias en la media y mediana de Cuarto con respecto a Quinto y Sexto, que por el contrario toman valores similares. La variabilidad es similar entre los tres años. Existen valores atípicos para Cuarto y Quinto año.



Mediante tests de normalidad de Shapiro-Wilk comprobamos que los datos no siguen una distribución normal, debido a que las pruebas arrojaron p valores inferiores a 0.05.

Ya que los datos no cumplen con los supuestos de normalidad, procedimos a realizar un test no paramétrico U de Kruskal-Wallis para contrastar si las diferentes muestras correspondientes a cada año están equidistribuidas.

H_0: Todas las muestras provienen de la misma distribución.

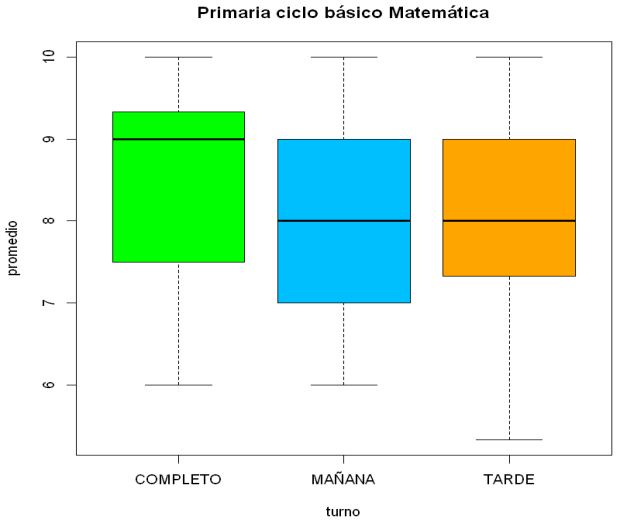
H_a: Al menos una de las muestras proviene de una distribución diferente.

Los resultados de la prueba realizada fueron Kruskal-Wallis chi-squared = 2.5378, df = 2, p-value = 0.2811. Con un valor p > 0.05 no se rechaza la hipótesis nula (H_0: Todas las muestras provienen de la misma distribución), es decir que nuevamente mediante la prueba realizada no se obtienen diferencias significativas. Se calcula la potencia del test ANOVA con el fin de estimar la potencia del test U de Kruskal-Wallis realizado obteniéndose una potencia de 0.524, Se estima que el test realizado (Kruskal-Wallis) posee una potencia menor.

A modo de cierre, tanto el análisis contrastando el ciclo básico con el ciclo orientado, como el análisis independiente de cada ciclo contrastando los diferentes cursos que los componen entre sí, podemos afirmar que mediante las pruebas realizadas, no hemos obtenido resultados que comprueben la existencia de diferencias significativas respecto a las matrículas en nuestro departamento.

Análisis del desempeño en Matemática en los diferentes turnos del ciclo básico de primaria.

Exploramos los datos de la muestra mediante boxplot y descriptivos, se observa en los valores promedios y de variabilidad un mismo comportamiento, solo el turno completo tiene una mediana mayor. la normalidad da valores muy bajos así que se utiliza test no paramétrico.



Turno	median	IQR	datos
COMPLETO	9	1.7	68
MAÑANA	8	2.0	860
TARDE	8	1.7	1028

El p valor en Kruskal.test dio: 0.003069 por lo tanto se rechaza la hipótesis nula. Al menos una muestra proviene de una población con una distribución distinta.

Se observan cómo se relacionan por separado las variables con el test pairwise.wilcox.test.

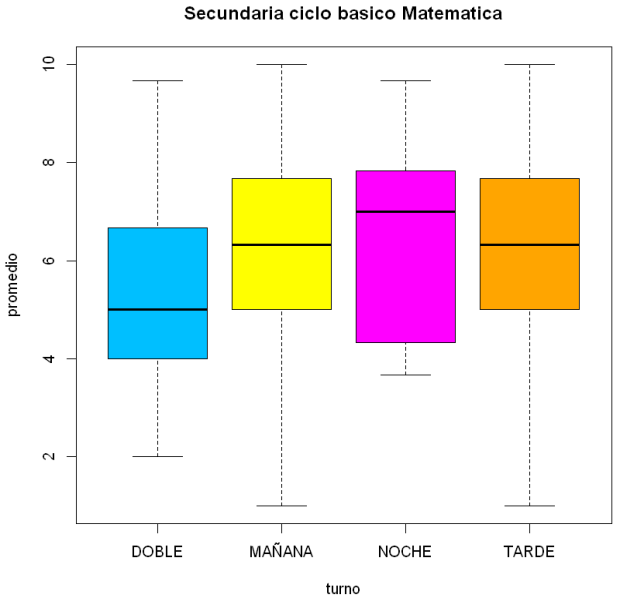
Se ve como existe poca relación en la distribución cuando se comparan a las variables entre Completo.

En conclusión en la escuela primaria ciclo básico turno Completo es el que tiene una distribución distinta en la materia matemática.

	COMPLETO	MAÑANA
MAÑANA	0.0023	-
TARDE	0.0023	0.9564

Análisis del desempeño en Matemática en los diferentes turnos del ciclo básico de secundaria.

Exploramos los datos de la muestra mediante boxplot y descriptivos, se observa en los valores de variabilidad un mismo comportamiento,excepto por el turno NOCHE que presenta mayor variabilidad. Al observar su medida de centralidad el turno DOBLE posee una menor mediana.



Turno	median	IQR	datos
DOBLE	5.0	2.7	165
MAÑANA	6.3	2.7	1877
NOCHE	7.0	3.4	40
TARDE	6.3	2.7	1016

La normalidad da valores muy bajos así que se utiliza test no paramétrico.

El p valor en Kruskal.test da 2.699e-10 y chi-squared 47.516 por lo tanto se rechaza la hipótesis nula. Al menos una muestra proviene de una población con una distribución distinta.

se observan cómo se relacionan por separado las variables con el test pairwise.wilcox.test

Se ve como existe poca relación en la distribución cuando se comparan a las variables entre DOBLE. En conclusión en la escuela secundaria ciclo básico el turno DOBLE es el que tiene una distribución distinta en la materia matemática.

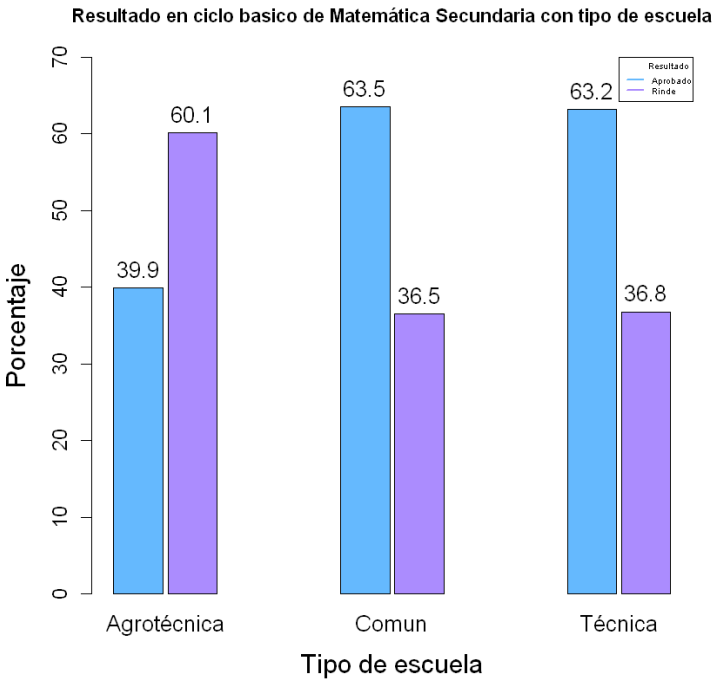
```
data:  basico$Promedio and basico$Turno

       DOBLE   MAÑANA NOCHE
MAÑANA 4.4e-11 -      -
NOCHE  0.062   1.000  -
TARDE  4.8e-10 1.000  1.000
```

ANÁLISIS BIVARIADO CON VARIABLES CUALITATIVAS

Análisis del Resultado en ciclo básico de Matemática Secundaria con tipo de escuela (Común- Técnica- Agrotécnica)

Al analizar el gráfico de la variable resultado en los diferentes tipos de escuelas, se observa que las escuelas de tipo Común y Técnica presentan un comportamiento similar en términos de distribución de notas. Sin embargo, se destaca que las escuelas Agrotécnicas muestran peores resultados, ya que tienen un mayor porcentaje de alumnos con bajo rendimiento y un menor porcentaje de alumnos aprobados.



En la siguiente imagen se puede observar una tabla de contingencia, tabla de porcentaje total, porcentaje fila, y porcentaje columna.

secundaria_matematica$Resultado	secundaria_matematica$ModEnsenanza Agrotécnica	Comun	Técnica	Row Total
Aprobado	73	1635	316	2024
	3.6%	80.8%	15.6%	62.1%
	39.9%	63.5%	63.2%	
	2.2%	50.2%	9.7%	
Rinde	110	941	184	1235
	8.9%	76.2%	14.9%	37.9%
	60.1%	36.5%	36.8%	
	3.4%	28.9%	5.6%	
Column Total	183	2576	500	3259
	5.6%	79.0%	15.3%	

A partir de los resultados obtenidos nos planteamos: ¿Existe dependencia de las notas de matemática según el tipo de escuela?

Al analizar los resultados que dio el test Pearson's Chi-squared podemos ver que el valor p extremadamente bajo sugiere que hay una asociación significativa entre las variables.

```
        Pearson's Chi-squared test

data:  secundaria_matematica$Resultado and secun
X-squared = 40.667, df = 2, p-value = 1.477e-09
```

```
          Agrotécnica Comun Técnica
Aprobado         -6.4   3.1     0.5
Rinde             6.4  -3.1    -0.5
```

observados

	Agrotécnica	Comun	Técnica
Aprobado	73	1635	316
Rinde	110	941	184

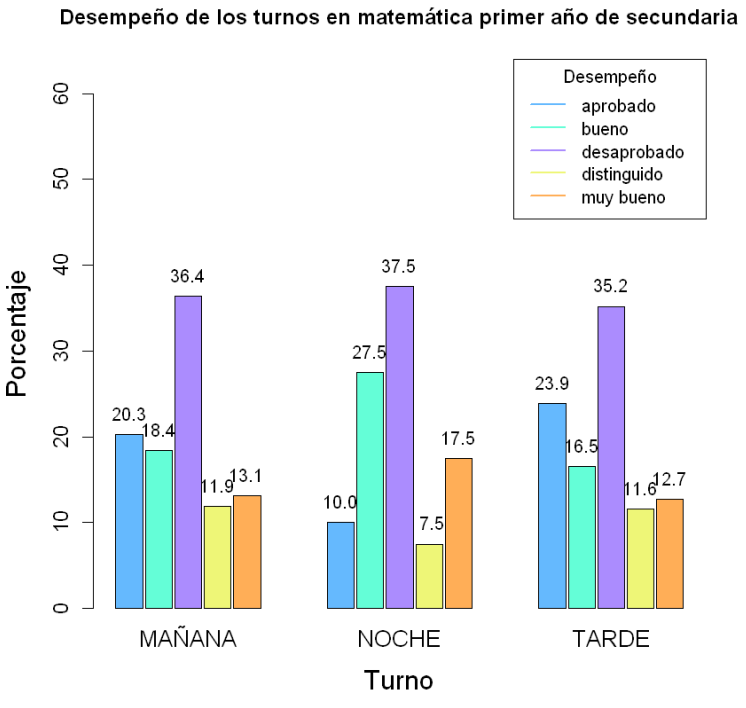
Esperados

	Agrotécnica	Comun	Técnica
Aprobado	113.65204	1599.8233	310.5247
Rinde	69.34796	976.1767	189.4753

Se procede a comparar los resultados observados y esperados. Podemos ver que existe una diferencia considerable en la escuela agrotécnica, en cambio la técnica hay poca diferencia.

Con CramerV validamos la tabla de contingencia y medimos la relación que existe entre las variables, su valor es 0.11170 y en ContCoef : 0.157000455. su relación es un poco baja pero considerando que el máximo es 0.6 decidimos seguir con el análisis.
Razones condicionales
escuelas agrotécnicas la razón de aprobar a no aprobar(rendir) es de 0.7 Entre las escuelas comunes y técnicas la razón de aprobar a no aprobar(rendir) es de 1.7
Razón de razones (odds ratio) o cociente de productos cruzados
El cociente entre razones condicionales simples permite ver la intensidad de la relación al comparar las diferentes categorías: La razón de aprobar a no aprobar(rendir) entre las escuelas agrotécnica y técnica es: 0.7/1.7=0.4. Entre agrotécnica y común de 0.7/1.7=0.4. Entre común y técnica de: 1.7/1.7=1.
Es decir de 100 estudiantes que aprueban en técnica o común solo 40 aprueban en la escuela agrotécnica.

El segundo análisis bivariado con materias cualitativas es en la secundaria ciclo básico, comparando los turnos por su desempeño



Al ver el gráfico podemos inferir de entrada como los 3 turnos se comportan de forma similar y como la variable “desaprobado” se diferencia del resto. Si bien podría resultar preocupante que la variable “desaprobado” sea la mayor, tiene un sentido lógico ya que las otras variables “aprobado” “bueno” “distinguido” y “muy bueno” se podrian pensar como parte de un todo el cual es “aprobados” y mirándolo desde esa perspectiva en los tres turnos existen muchos más aprobados que desaprobados.

```
        Pearson's Chi-squared test

data:  x$Desempeño and x$Turno
X-squared = 11.487, df = 8, p-value = 0.1756
```

Al analizar los resultados que dio el test Pearson's Chi-squared podemos ver que el valor p es alto por esto, no se rechaza la hipótesis nula, y si concluye que existe independencia entre los turnos mañana, noche y tarde al compararlos con la variable desempeño.

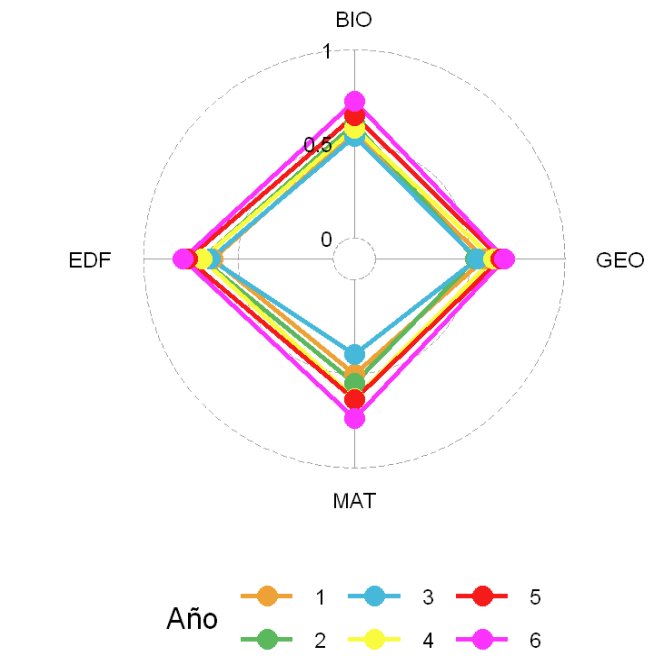
ANÁLISIS MULTIVARIADO

Comparación de los promedios de los diferentes años de cursado en las materias troncales de Secundaria

Se puede observar que el valor más bajo se encuentra en el tercer año con la materia de matemática. Este hecho también se refleja en el gráfico radial, donde se puede apreciar que la variable correspondiente al tercer año (representada en azul) se encuentra más cerca del centro. Además, se puede notar que el sexto año presenta los valores más altos, siendo la forma geométrica asociada a este año la más grande dentro del gráfico radial.

Vector de medianas

	BIO	GEO	MAT	EDF
1	6.00	7.00	5.33	6.67
2	6.67	6.33	6.00	7.33
3	5.33	6.00	4.50	6.67
4	6.00	7.33	7.33	7.50
5	7.00	7.67	7.00	8.33
6	8.00	7.67	8.00	8.33



Se calcula la matriz de varianza y covarianza en las materias troncales de los distintos años para describir la dispersión y la relación lineal entre las variables. Al analizar la covarianza, se observa que todos los valores son cercanos a cero, lo cual indica una relación débil o nula entre las variables en términos de variación conjunta.
En cuanto a la varianza, se encuentra que la mayor dispersión se da en la materia de matemática en los diferentes años, mientras que la educación física presenta menor dispersión. Además, se destaca que el tercer año muestra la mayor variabilidad en las materias.

1ro

	BIO	GEO	MAT	EDF
BIO	0.98	0.54	0.91	0.49
GEO	0.54	0.95	0.73	0.26
MAT	0.91	0.73	1.27	0.62
EDF	0.49	0.26	0.62	0.83

2do

	BIO	GEO	MAT	EDF
BIO	1.04	0.74	0.81	0.24
GEO	0.74	1.12	0.86	0.42
MAT	0.81	0.86	1.46	0.41
EDF	0.24	0.42	0.41	0.74

3ro

	BIO	GEO	MAT	EDF
BIO	2.36	1.50	2.61	1.09
GEO	1.50	1.88	2.10	0.89
MAT	2.61	2.10	5.81	2.31
EDF	1.09	0.89	2.31	1.89

4to

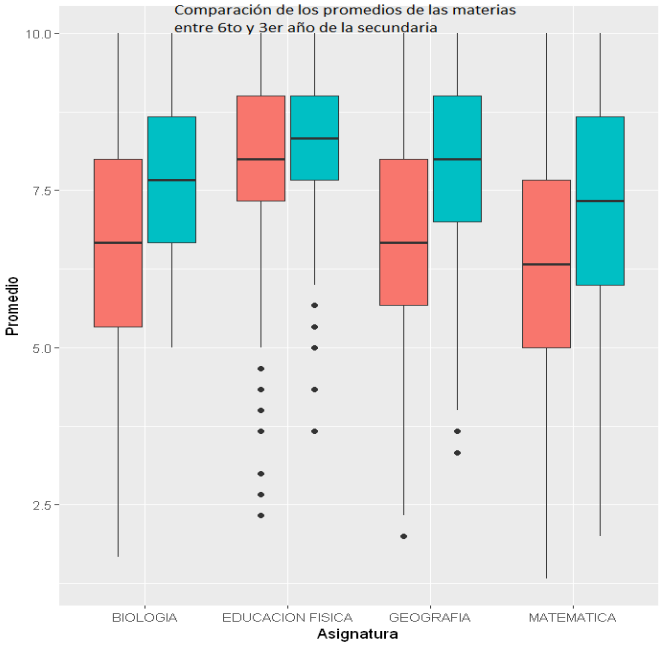
	BIO	GEO	MAT	EDF
BIO	2.14	0.82	1.41	0.30
GEO	0.82	2.40	1.73	-0.19
MAT	1.41	1.73	3.21	0.31
EDF	0.30	-0.19	0.31	1.36

5to

	BIO	GEO	MAT	EDF
BIO	2.14	0.82	1.41	0.30
GEO	0.82	2.40	1.73	-0.19
MAT	1.41	1.73	3.21	0.31
EDF	0.30	-0.19	0.31	1.36

6to

	BIO	GEO	MAT	EDF
BIO	2.14	0.82	1.41	0.30
GEO	0.82	2.40	1.73	-0.19
MAT	1.41	1.73	3.21	0.31
EDF	0.30	-0.19	0.31	1.36



Al observar la diferencia que existía entre sexto año y tercer año decidimos realizar el test Wilcoxon para ver si existe una diferencia significativa entre ellos.

```
        Wilcoxon rank sum test with continuity correction

data:  variable1 and variable2
W = 2261274, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Hipótesis nula (H0): No hay diferencia significativa en los promedios de las materias entre 6to y 3er año de la secundaria.

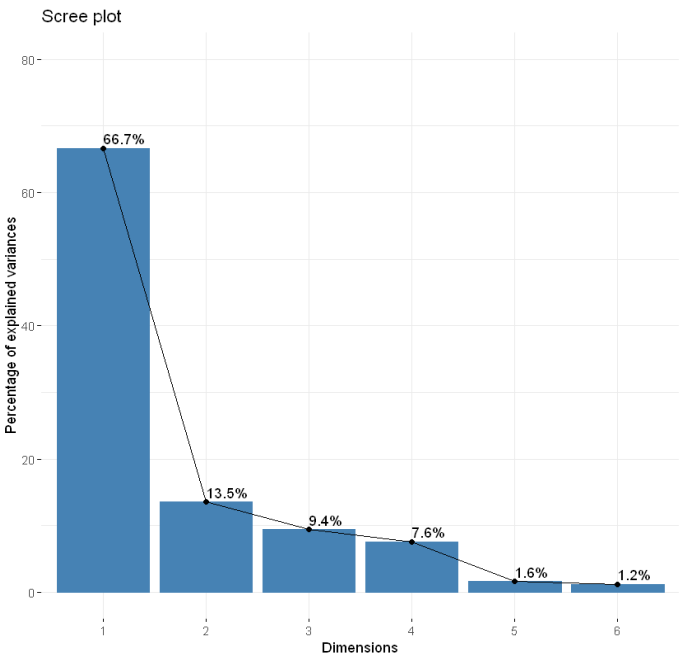
Hipótesis alternativa (H1): Existe una diferencia significativa en los promedios de las materias entre 6to y 3er año de la secundaria.

Al ver que el test wilcoxon devuelve un p valor menor a 0.05 se rechaza la hipótesis nula, si existe una diferencia significativa en los promedios de las materias entre 6to y 3er año de la secundaria

MÉTODOS DE REDUCCIÓN DE DIMENSIONALIDAD

Al realizar la prueba "det" en nuestra matriz de correlaciones, obtuvimos un valor muy cercano a cero (0.005757152), lo que indica un alto nivel de colinealidad entre las variables involucradas en la matriz.

El KMO (Kaiser-Meyer-Olkin) o los valores MSA (Medida de Muestreo de Adecuación) indican una adecuación moderada de los datos para el análisis factorial. Esto significa que podemos llevar a cabo análisis factoriales con cierta confianza, ya que los resultados de MSA arrojaron valores cercanos a 1 (MSA total = 0.8).

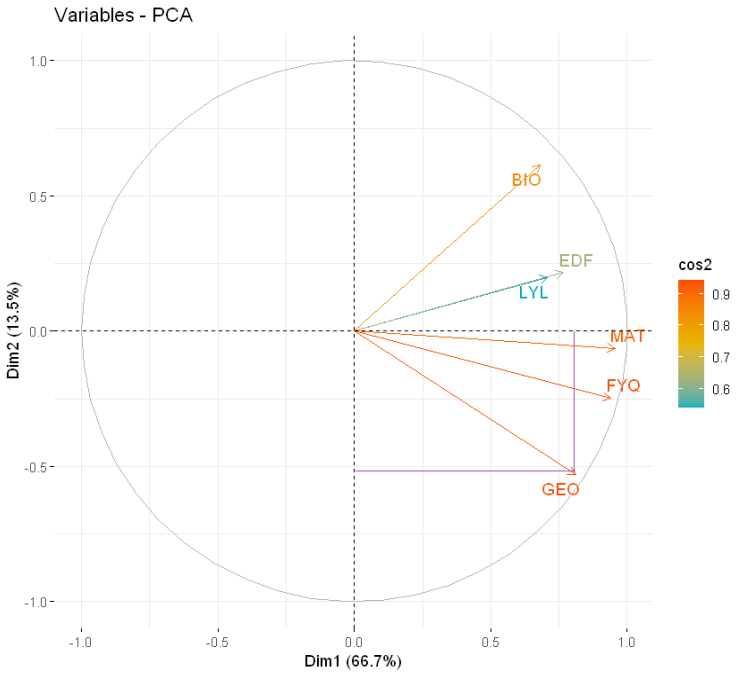


Resultados del Análisis de Componentes Principales (ACP)

Los dos primeros factores explican aproximadamente el 80% de la varianza de las variables originales, lo que indica la solidez de nuestro modelo factorial. Además, podemos observar las proporciones de varianza explicadas por

cada uno de ellos. En este sentido, el primer factor explica más del 66% de la varianza, mientras que el segundo factor explica alrededor del 13%.

En general, todas las variables muestran una correlación positiva, ya que se agrupan juntas o están próximas entre sí. Específicamente, se destaca la dependencia entre lengua y literatura (LYL) y educación física (EDF) . No se encontraron variables que presenten una correlación negativa.



Entre más cerca del eje x están las variables más aporta a la componente principal 1, y entre más cerca del eje Y mayor aporte a la componente 2. lo que podemos observar en el gráfico es que en general las variables aportan a la componente 1

A modo de conclusión, el análisis realizado proporcionó información relevante sobre el perfil y desempeño de los estudiantes en el sistema educativo. Los hallazgos principales son los siguientes:

Nivel educativo: El nivel secundario cuenta con una mayor cantidad de estudiantes la cual es 5733, en comparación con el nivel primario, que tiene 4569.

Sector público: El sector público muestra una alta prevalencia en la población estudiantil, con un 90% de los estudiantes perteneciendo a este sector.

Turno escolar: Se observa una mayor proporción de estudiantes en el turno mañana (61%) en comparación con el turno tarde (39%).

Distribución por año: En el nivel primario, la distribución de estudiantes por año es homogénea. Sin embargo, en el nivel secundario, se encontraron diferencias significativas en la distribución, indicando una desigualdad en la cantidad de estudiantes por año.

Desempeño en materias: En el nivel secundario, se identificó que la materia de matemática presentó el peor desempeño, con un porcentaje de aprobados del 58% en los dos primeros años. En segundo lugar se encuentra lengua y literatura, también en los primeros dos años, con un porcentaje de aprobados del 64%.

Correlación entre matemática y lengua: Se observó una correlación positiva entre matemática y lengua en el nivel secundario, con un valor de 0.76. A medida que aumenta el año de cursado, la

correlación disminuye ligeramente a 0.57. En el nivel primario, se encontró una correlación más fuerte, con valores de 0.94 y 0.88, respectivamente.

Correlación entre matemática y educación física: Se encontró una correlación positiva entre matemática y educación física, con un valor de rho = 0.42 y un valor p muy significativo (< 2.2e-16), lo que indica una asociación estadísticamente significativa entre ambas asignaturas.

Mejoría en educación física: Se destaca que un 21% de los estudiantes que inicialmente no aprobaron el primer y segundo trimestre de educación física lograron aprobar finalmente la materia, lo que indica una mejora significativa durante el período.

Estos hallazgos ofrecen información valiosa para comprender el panorama educativo, identificar áreas de mejora y tomar decisiones informadas para el desarrollo y apoyo de los estudiantes en su trayectoria educativa.

Comparación de medias:
En el departamento las notas de secundaria del tercer trimestre son mayores a las del primero.

Los promedios de Lengua del 50% de los estudiantes de secundaria estuvo por encima de 6.

No se encontraron diferencias significativas en el ciclo básico al comparar el primer, segundo y tercer año, con una potencia estadística menor a 0.524. Del mismo modo, no se encontraron diferencias significativas en el ciclo orientado al comparar el cuarto, quinto y sexto año, con una potencia estadística menor a 0.524.

En la escuela primaria ciclo básico, el turno COMPLETO muestra una distribución de notas en matemática diferente a los turnos mañana y tarde.
En la escuela secundaria ciclo básico, el turno DOBLE muestra una distribución de notas en matemática diferente a los turnos tarde, mañana y noche.

Análisis bivariado con variables cualitativas:

En la materia de matemática, las escuelas agrotécnicas tienen un desempeño inferior en comparación con las escuelas comunes y técnicas, mientras que estas últimas tienen un desempeño similar. Se encontró una asociación significativa entre las variables. De 100 estudiantes que aprueban en técnica o común solo 40 aprueban en la escuela agrotécnica..
Se observa independencia entre los turnos mañana, noche y tarde al compararlos con el desempeño en el ciclo básico de secundaria.

Análisis multivariado:

En cuanto a la covarianza, se encontró que las variables tienen una relación débil o nula en términos de variación conjunta, ya que todos los valores están cercanos a cero.
En relación a la varianza, se observa que la mayor dispersión se encuentra en la materia de matemática en los diferentes años, mientras que la educación física presenta menor dispersión.
El sexto año muestra las mejores notas, mientras que el tercer año presenta las peores, y estas diferencias son significativas.

Métodos de reducción de dimensionalidad:

Los dos primeros factores explican aproximadamente el 80% de la varianza de las variables originales, lo que indica la solidez de nuestro modelo factorial.
Todas las variables muestran una correlación positiva, ya que se agrupan o están cercanas entre sí. Específicamente, se destaca la dependencia entre lengua y literatura (LYL) y educación física (EDF).