UNER Facultad de Ingeniería

Año: 2023

Carrera: Tecnicatura Universitaria en Procesamiento y Explotación de Datos

Cátedra: Exploración de datos multivariados

Tema: Trabajo Práctico N° 1

Profesor/a: Melisa Fernandez

Alumno/s: Armú Yamil, Zamora Rodrigo

El presente trabajo es el resultado del análisis llevado a cabo sobre las escuelas del departamento de Colón.

Con los datos disponibles se realizaron técnicas de limpieza y organización, estudios de representatividad de la muestra respecto a la provincia, 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 gráficos.

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
- Google Workspace (Meet, Drive, Docs, etc)

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.

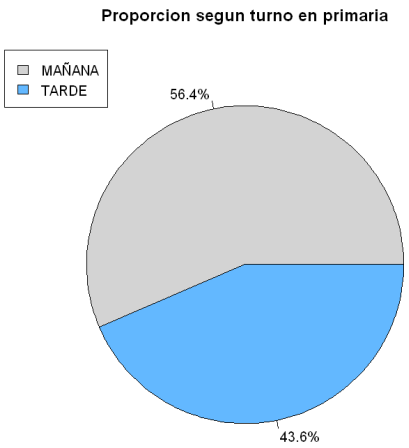
Posteriormente se realizó un estudio de representatividad de los datos del departamento con respecto a la provincia, se obtuvieron los siguientes resultados.

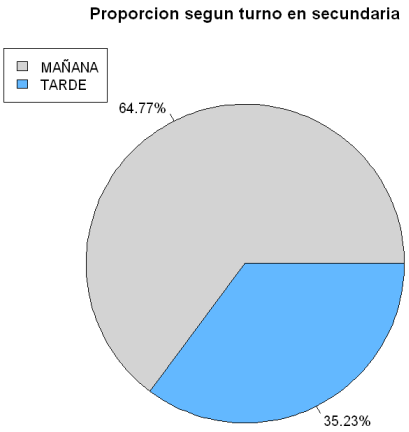
PRIMARIA	
Cantidad de escuelas	1266
Cantidad de escuelas cargadas	25
Matrícula total:	142480
Matrícula cargada en notas:	4596
Representatividad	384

SECUNDARIA							
Cantidad de escuelas: 572			Cantidad de escuelas cargadas: 25				
	1°	2°	3°	4°	5°	6°	7°
Matrícula total	25317	25989	22569	23430	19739	17649	2226
Matrícula cargada en notas*	1262	1244	955	882	698	653	99
Representatividad	379	380	379	379	378	377	329

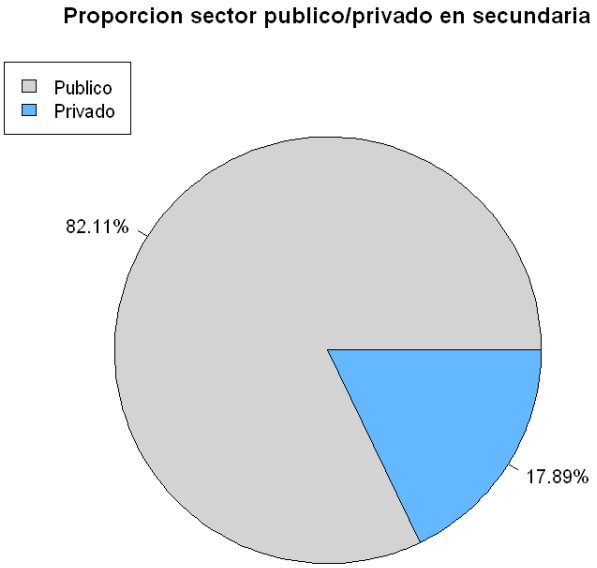
Dónde podemos afirmar que los datos cumplen con el tamaño mínimo muestral.

Por otra parte, se realizó un análisis sobre la distribución de datos según modalidad (primaria/secundaria), turno (mañana/tarde) y sector (público/privado).



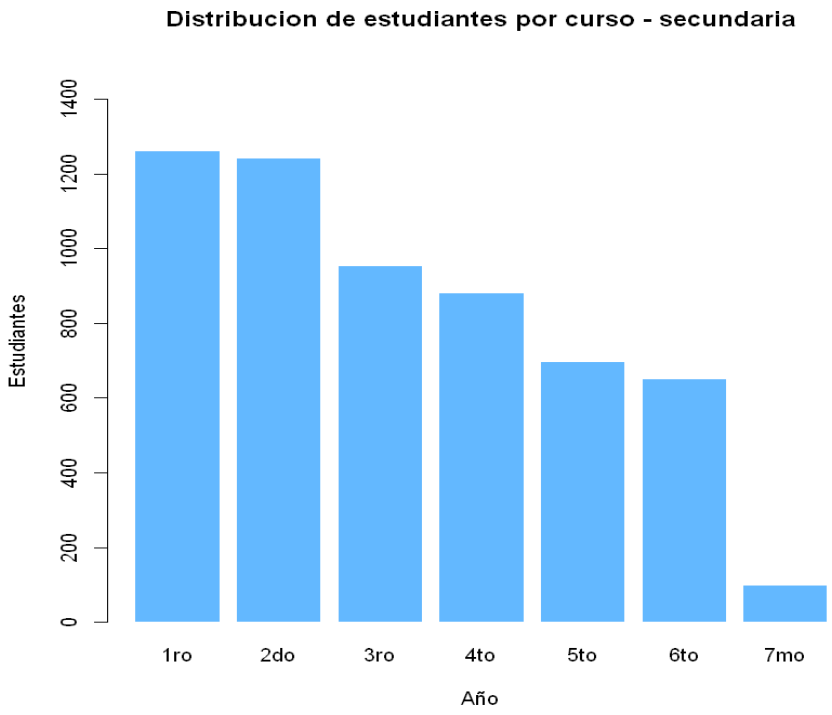


Donde se observó un mayor número de estudiantes en secundaria con respecto a primaria. Prevalencia en el turno mañana con el 56.4% para primaria y 65.7% en secundaria.

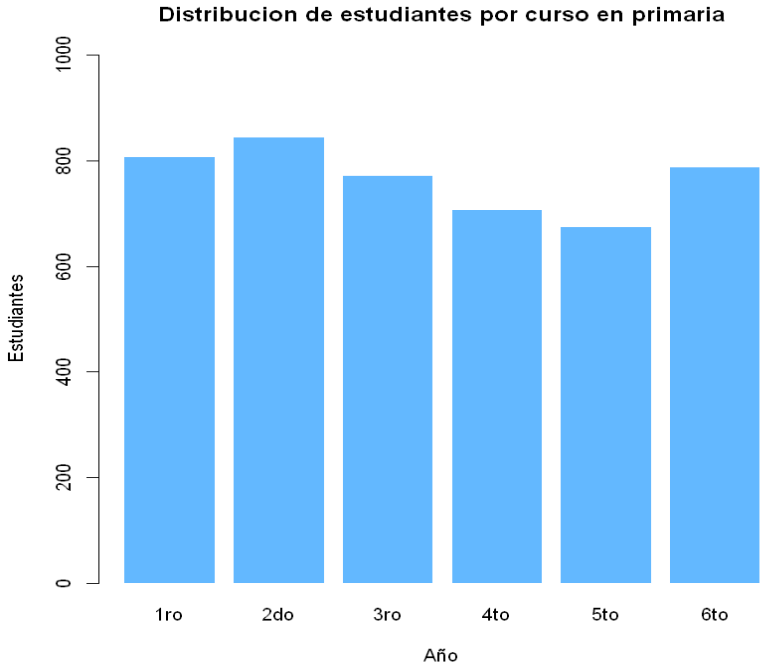


Mayor cantidad perteneciente al sector público con el 100% para primaria y el 90% en el caso de secundaria.

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



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



Por el contrario, para el conjunto de primaria observamos una distribución homogénea, siendo segundo año el de mayor matrícula.

Se obtienen medidas de centralidad y dispersión correspondientes a los promedios de materias como matemática y lengua.

Matemática - Secundaria

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 - Secundaria

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

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 al menos el 50% de los estudiantes de todos los años tuvieron promedio mayor a 6.

En términos de la desviación estándar, podemos observar que la variabilidad en los resultados 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, podemos concluir que 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.

muestra de 1179 (lengua) y 1231 (matemática)

medidas de centralidad:

matemática media 6.5 moda 6 y mediana 6.33

lengua y literatura: media 6.80 moda 6 mediana 6.67

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

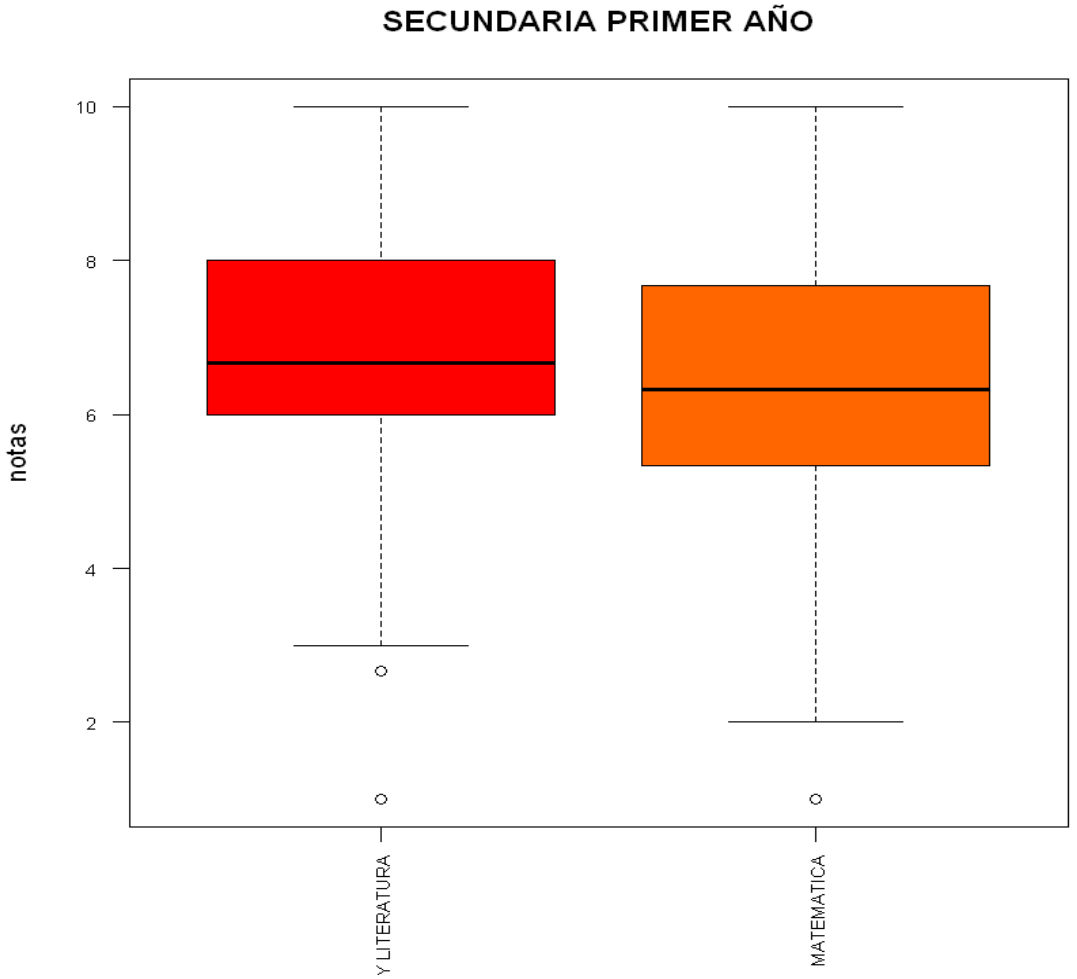
la media, es representativa para ambas ya que las variables casi no tienen valores extremos y la media comparte valores similares con la mediana y moda.

Se realiza lillie.test en las dos materias y su p valor da un número muy chico menor a 0.05 por lo tanto no tienen una distribución normal.

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.

rango intercuartil matemática: 2.3 lengua: 2

El tercer cuartil o 75% de las notas en matemática es 7.6 y lengua 8



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

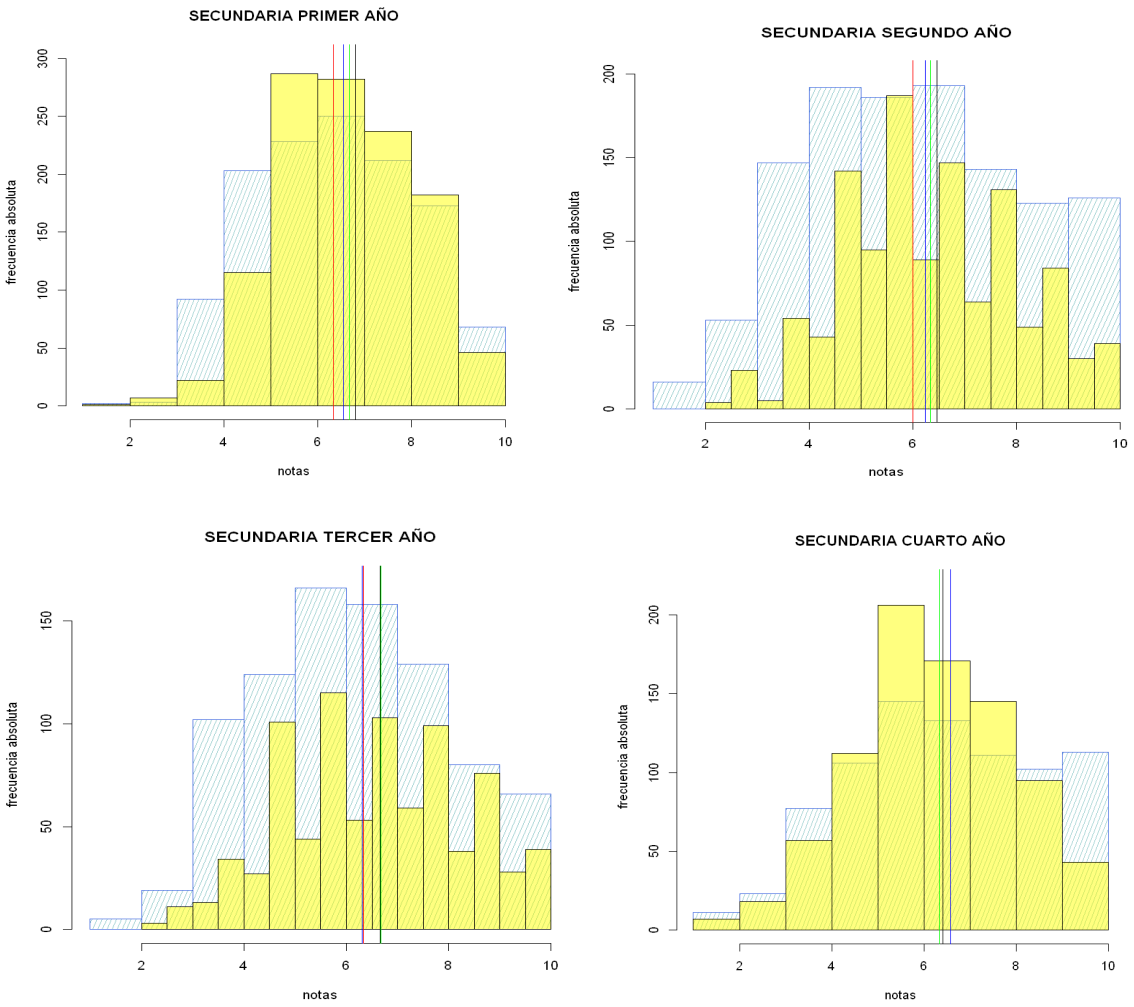
las materias tienen una dispersión normal, matemática al tener mayor desvío estándar y un mayor rango intercuartil (matemática: 2.3 lengua: 2) tiene una mayor variabilidad.

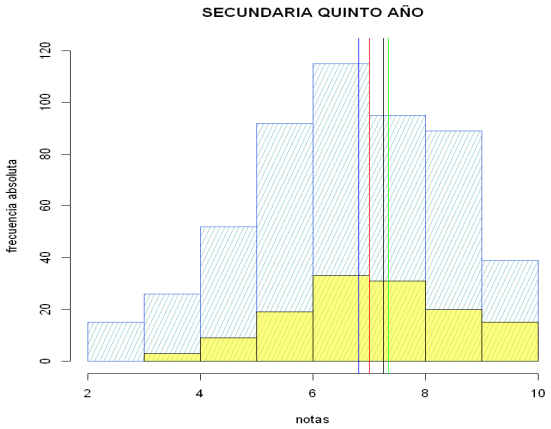
Existen dos outliers en lengua y literatura y uno de matemática con notas muy bajas menores a 3 en una muestra de 1179 (lengua) y 1231 (matemática) lo cual es poco para una muestra tan grande.

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.

Relación 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.

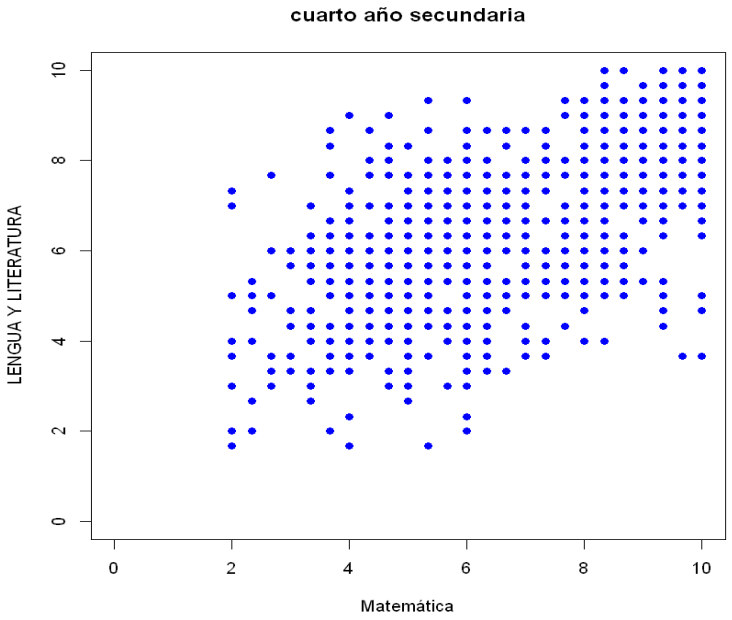
Utilizamos el método de correlación de Spearman para realizar una prueba de hipótesis para evaluar si hay una asociación monotónica significativa (positiva o negativa) entre las dos variables.

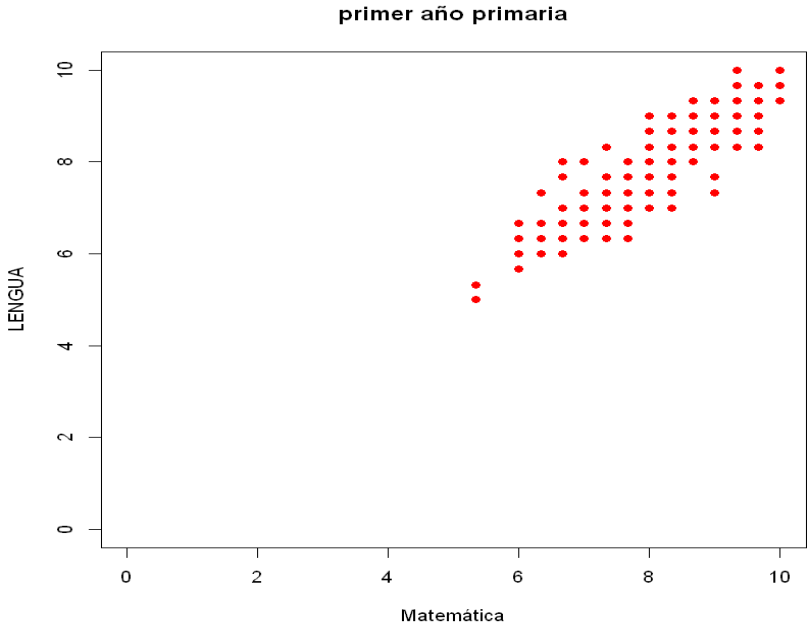
No se encontró alumnos que tengan notas de matemática y lengua y literatura en quinto año de secundaria.

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

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 66.4% hasta el 92.2%. La materia de Lengua y Literatura también presenta una variabilidad significativa en sus resultados.

están ordenadas de 1 a 6

Asignatura	Resultado
BIOLOGIA	0.6893543
EDUCACION FISICA	0.9294643
FISICOQUIMICA	0.7212066
GEOGRAFIA	0.7147651
HISTORIA	0.7454545
INGLES	0.7733333
LENGUA Y LITERATURA	0.6424958
MATEMATICA	0.5810008

Asignatura	Resultado
BIOLOGIA	0.6893543
EDUCACION FISICA	0.9294643
FISICOQUIMICA	0.7212066
GEOGRAFIA	0.7147651
HISTORIA	0.7454545
INGLES	0.7733333
LENGUA Y LITERATURA	0.6424958
MATEMATICA	0.5810008

Asignatura	Resultado
BIOLOGIA	0.6639535
EDUCACION FISICA	0.9446328
FISICOQUIMICA	0.7514654
GEOGRAFIA	0.7049724
HISTORIA	0.7222222
INGLES	0.7333333
LENGUA Y LITERATURA	0.6809015
MATEMATICA	0.6124853

Asignatura	Resultado
BIOLOGIA	0.7135741
EDUCACION FISICA	0.9552600
FISICA	0.7250946
GEOGRAFIA	0.7367803
HISTORIA	0.8285714
INGLES	0.6666667
LENGUA Y LITERATURA	0.6510539
MATEMATICA	0.6333739
QUIMICA	0.6971154

Asignatura	Resultado
BIOLOGIA	0.7916667
EDUCACION FISICA	0.9737991
FISICA	0.8049242
GEOGRAFIA	0.8678161
INGLES	0.7982456
LENGUA Y LITERATURA	0.8307692
MATEMATICA	0.7552581
QUIMICA	0.8057785

Asignatura	Resultado
BIOLOGIA	0.9215686
EDUCACION FISICA	0.9777778
FISICA	0.9419355
GEOGRAFIA	0.9394531
MATEMATICA	0.8045541
QUIMICA	0.8962963

Análisis de los alumnos que desaprobaron el primer y segundo trimestre, pero de todas formas pudieron aprobar la materia.

porcentaje de alumnos que pudieron aprobar la materia habiendo desaprobado primer y segundo trimestre:

SECUNDARIA

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

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.

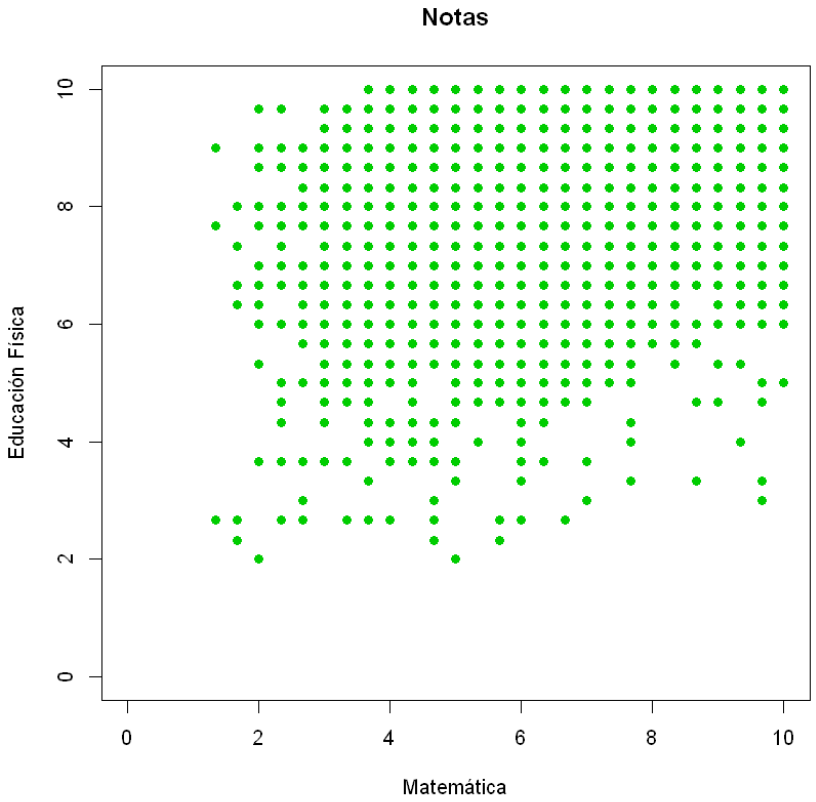
En primer lugar se comprobó que ambas variables tengan una cantidad similar de datos, y se crearon histogramas que permitieron visualizar las respectivas distribuciones.
Se realizaron pruebas de normalidad, que, debido a la cantidad de datos para cada variable, de 5155 y 5369, para matemática y ed. física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

H_0 : La variable sigue una distribución normal.

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

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 = 0.3343801 y p-value < 2.2e-16.



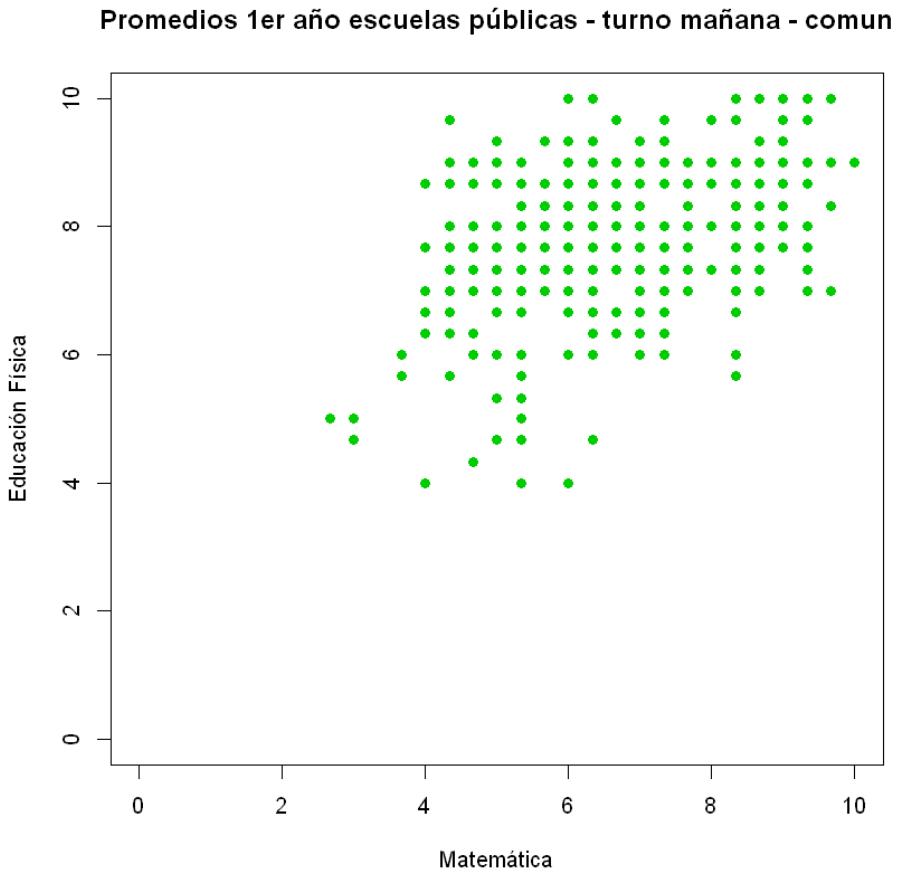
Se repite el análisis, ésta vez se aplicaron otros filtros sobre los datos (turno mañana, modalidad común, escuela pública), a fin de eliminar variables que podrían estar interfiriendo en los resultados esperados.

H_0 : La variable sigue una distribución normal.

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

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, se observa 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.


Promedios 1er año escuelas públicas - turno mañana - comun
Educación Física
Matemática
0
2
4
6
8
10

Conclusiones.

Mayor cantidad de estudiantes en secundaria (5733), en comparación a primaria (4569).

Prevalencia del sector público, 100% en primaria y 90% en secundaria.

Mayor proporción en turno mañana (61%) frente al turno tarde (39%).

Distribución por año homogénea en el caso de primaria, desigual para secundaria.

La materia con peor desempeño en secundaria fue matemática, con un porcentaje de aprobados del 58% para los dos primeros años. Seguido por lengua y literatura, siendo también primero y segundo los años con porcentaje de aprobados más bajo 64%.

Existe una correlación positiva entre matemática y lengua, en secundaria con un valor de 0.76, que decae a 0.57 con respecto aumenta el año de cursado. El mismo comportamiento se observó en primaria pero con valores mayores, de 0.94 a 0.88.

Existe una correlación positiva entre matemática y educación física, con un valor rho = 0.42 y p-value < 2.2e-16.

Un 21% de los estudiantes que desaprobaron el primer y segundo trimestre de educación física lograron aprobar finalmente la materia.