



UNIVERSIDAD NACIONAL DE ENTRE RÍOS

FACULTAD DE INGENIERÍA

**CARRERA: Tecnicatura Universitaria en Procesamiento y
Explotación de Datos**

MATERIA: Exploración de datos multivariados

Nombre de la Actividad: Trabajo Práctico parte 1

**Tema: Primer informe de las notas de las escuelas del
departamento La Paz**

Fecha de Entrega: 25/04/2023

Profesores: Esp. Prof. Melisa Fernández

Alumnos: Venturini, Angelo

Ruiz Diaz, Enzo

Primer informe de las notas de las escuelas del departamento La Paz

ÍNDICE:

Introducción	3
Desarrollo	4
Exploración	4
Reorganización y limpieza	9
Análisis de las características generales de la educación	10
Rendimiento académico de matemática y lengua segmentado por año académico.	16
Primaria: matemática	17
Primaria: Lengua	18
Secundaria: matemática	19
Secundaria: Lengua y literatura	19
Correlaciones	22
Conclusiones	26
Anexo	28

Introducción

En el presente informe encontrarán un análisis de una base de datos que contiene las escuelas primarias y secundarias de educación común y técnica del departamento La Paz de la provincia de Entre Ríos. El análisis tiene el fin de diagnosticar el estado general de la educación en dicho departamento, y en particular analizar el rendimiento de los estudiantes en las materias básicas de secundaria (Matemática, Lengua, Geografía, Historia, Biología, Física, Química, Inglés, Educación Física) y primaria cursadas en el año 2022, con la idea de asesorar en la implementación de programas complementarios.

Se realizó el análisis con la herramienta Jupyter Notebook y el lenguaje de programación R. Además para analizar la representatividad se utilizaron dos bases de datos que contienen la totalidad de escuelas primarias y secundarias de la provincia. Dichas bases de datos fueron proporcionadas por la cátedra.

Desarrollo

Exploración

En una primera instancia el grupo realizó una exploración del dataset donde se reconoció la dimensión del mismo, se analizaron los nombres y tipo de las variables, como también se analizaron los tipos de los datos cargados. Por otra parte se verificó la existencia de datos ausentes, datos duplicados y datos mal cargados. En otra etapa de la exploración, analizamos si había variables redundantes o que no aporten información al problema. En la tabla 1 se muestra la información obtenida del proceso de exploración.

Tabla 1

Análisis exploratorio

Departamento:	La Paz
Tamaño	12127150
N.º de filas	485086
N.º de columnas	25
Variables Cuantitativas	idalumno, documento, anioLectivo, nota, idSuborganizacion, idDivision, idorganizaciones,
Variables Cualitativas	CUE, orden, nivel, EsPrivada, turno, AñoCursado, periodoEvaluatorio, asignatura, NivelEnsenanza, ddivision, esMultianio, Modalidad, ModEnsenaza
Variables a modificar nombre	anioLectivo, NivelEnsenaza, ModEnsenaza
Variables mal declaradas	curso, turno, Modalidad, nota, EsPrivada, ModEnsenansa
Variables redundantes	idalumno
Variables que no aportan información	idSuborganizacion, idDivision, idorganizaciones, esMultianio, motivo_ausente
Variables con NA	motivo_ausente, observaciones
Variables con datos sucios	AñoCursado, asignatura
Variables importantes	documento, nota, periodoEvaluatorio, asignatura

Además, se realizó un conteo de escuelas y de matrículas donde obtuvimos para escuelas primarias, 49 escuelas y 5151 matrículas luego estudiamos la representatividad de la muestra, donde utilizando una calculadora de tamaño de muestra (SurveyMonkey) para una población de 56 escuelas primarias, información obtenida del dataset común proporcionado por la cátedra, con un nivel de confianza del 95% se necesita un tamaño de muestra de 49 y para una población de 5987 matrículas con un nivel de confianza del 95% se necesita un tamaño de muestra de 362. Contrastando esta información con nuestra muestra concluimos que es representativa.

Para escuelas secundarias encontramos 22 escuelas y 4564 matrículas.

Tabla 2

Estudio de la representatividad de la muestra de escuelas secundarias

SECUNDARIA							
Cantidad de escuelas: 22				Cantidad de escuelas cargadas: 22			
	1°	2°	3°	4°	5°	6°	7°
Matrícula total	1067	1078	1000	1036	885	819	101
Matrículas cargadas	892	844	849	763	598	561	57
Tamaño mínimo de muestra	283	284	278	281	269	262	81

Como podemos ver en la tabla 2, la muestra para séptimo año no es representativa por lo que no tendremos en cuenta este año para los posteriores análisis.

También realizamos un resumen de los estadísticos para todos los cursos de primaria de las materias matemática y lengua.

1°	conteo	media	varianza	desvío	min	max	IQR	rango	mediana
Lengua	985	8,36	1,29	1,14	6	10	1,67	4	8,67
Matemática	1010	8,47	1,25	1,12	6	10	1,67	4	8,67

2°	conteo	media	varianza	desvío	min	max	IQR	rango	mediana
Lengua	862	8,5	1,30	1,14	6	10	1,67	4	8,67
Matemática	830	8,49	1,35	1,16	6	10	1,67	4	8,67

3°	conteo	media	varianza	desvío	min	max	IQR	rango	mediana
Lengua	867	8,53	1,15	1,07	6	10	1,67	4	8,67
Matemática	866	8,56	1,29	1,14	6	10	2	4	8,67

4°	conteo	media	varianza	desvío	min	max	IQR	rango	mediana
Lengua	856	8,27	1,11	1,05	6	10	1,33	4	8,33
Matemática	851	8,31	0,994	0,997	6	10	1,33	4	8,33

5°	conteo	media	varianza	desvío	min	max	IQR	rango	mediana
Lengua	824	8,16	1,13	1,06	6	10	1,67	4	8,33
Matemática	820	8,16	1,23	1,11	6	10	1,67	4	8,33

6°	conteo	media	varianza	desvío	min	max	IQR	rango	mediana
Lengua	838	8,09	1,14	1,07	6	10	1,67	4	8
Matemática	835	8,12	1,24	1,11	6	10	1,67	4	8

Ahora lo realizamos para los años de las escuelas secundarias

1°	conteo	media	varianza	desvío	min	max	IQR	rango	mediana
Lengua y Literatura	850	6,95	2,59	1,61	2,33	10	2,33	7,67	7
Matemática	887	6,8	2,98	1,73	1	10	2	9	6,67

2°	conteo	media	varianza	desvío	min	max	IQR	rango	mediana
Lengua y Literatura	756	6,34	3,93	1,98	1	10	2,67	9	6,5
Matemática	843	6,32	3,69	1,92	1,67	10	3	8,33	6,33

3°	conteo	media	varianza	desvío	min	max	IQR	rango	mediana
Lengua y Literatura	918	6,51	3,40	1,84	1	10	2,5	9	6,33
Matemática	936	6,23	3,93	1,98	1	10	2,67	9	6,33

4°	conteo	media	varianza	desvío	min	max	IQR	rango	mediana
Lengua y Literatura	1377	6,04	3,45	1,86	1	10	2,67	9	6
Matemática	1414	6,14	4,10	2,03	1	10	3	9	6,33

5°	conteo	media	varianza	desvío	min	max	IQR	rango	mediana
Lengua y Literatura	131	7,52	2,14	1,46	2,33	10	2,33	7,67	7,67
Matemática	1548	6,63	3,10	1,76	1	10	2,67	8,33	6,67

6°	conteo	media	varianza	desvío	min	max	IQR	rango	mediana
Lengua y Literatura	9	8,30	0,457	0,676	7	9	1	2	8
Matemática	1451	6,91	3,30	1,82	1	10	2,67	9	7

Reorganización y limpieza

Luego de la exploración de la base de datos se concluye que el estado general de la misma es mala y no se podría realizar el análisis encargado sin antes realizar ciertas modificaciones. Estas incluyen renombrar variables para que se entienda la información que estas contienen, cambiar el formato de los valores para poder procesar dichos datos, normalizar ciertos datos de tipo string o char para que haya homogeneidad y agregar variables nuevas que ayuden al análisis como es el caso de la variable promedio y curso.

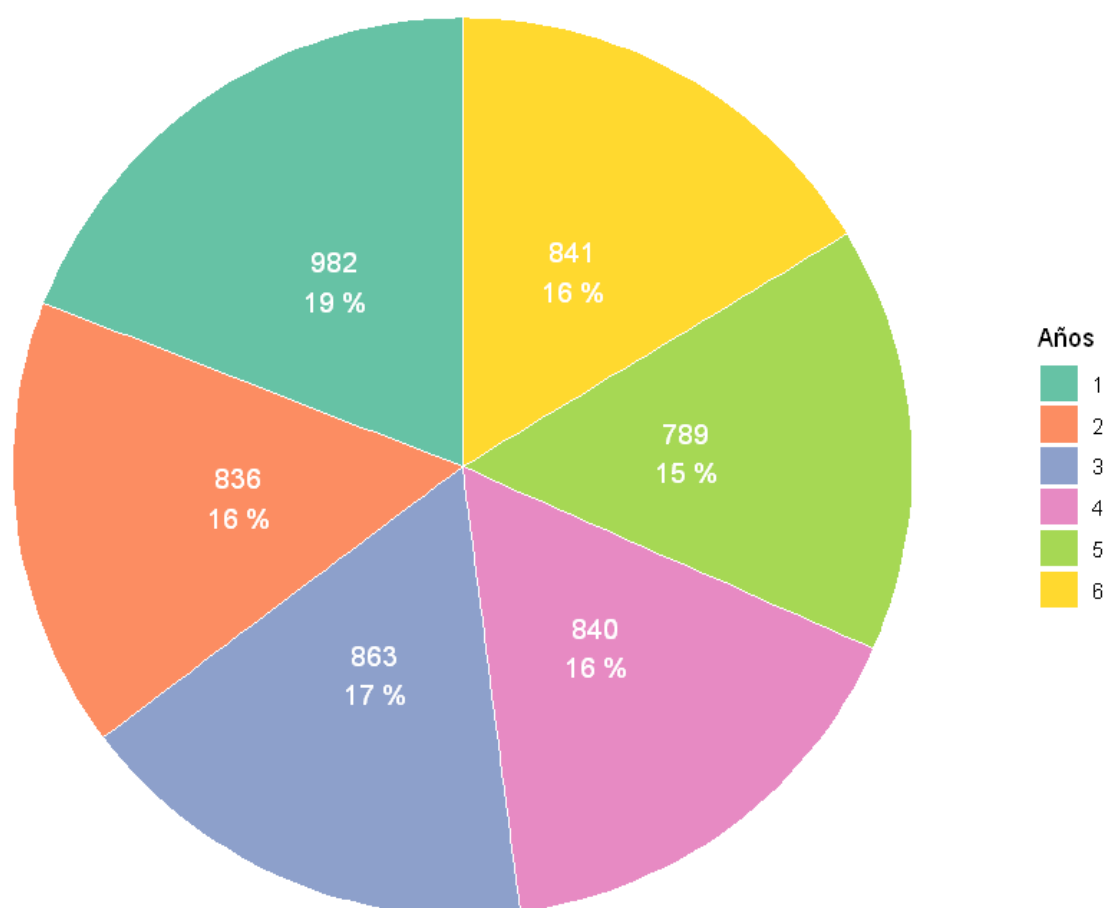
Encontramos que la columna nivel enseñanza está relacionada con la columna nivel, con la diferencia de que esta última nos brinda menos información por lo que decidimos eliminarla. También, eliminamos la columna de año ya que todos nuestros datos son del año 2022 y sería redundante dejarla. Adicionalmente, solo se dejaron las filas cuyos promedios pudieron ser calculados con las principales materias troncales de cada nivel de enseñanza, siendo para primaria ciencias naturales, ciencias sociales, matemática, lengua, educación física e inglés. Para secundaria, matemática, lengua y literatura, educación física, geografía, historia, biología, físico-química, química, física e inglés.

Análisis de las características generales de la educación

Primeramente analizamos la distribución de la matrícula, segmentada por año, de las escuelas primarias. Como se observa en el gráfico 1, el mayor número de matriculados pertenecen al segmento de primer año con 19 %, seguido de tercer grado con 17%,

Gráfico 1

Distribucion Porcentual de la Matricula de Primaria de 2022 en el Departamento La Paz por Año de Cursado

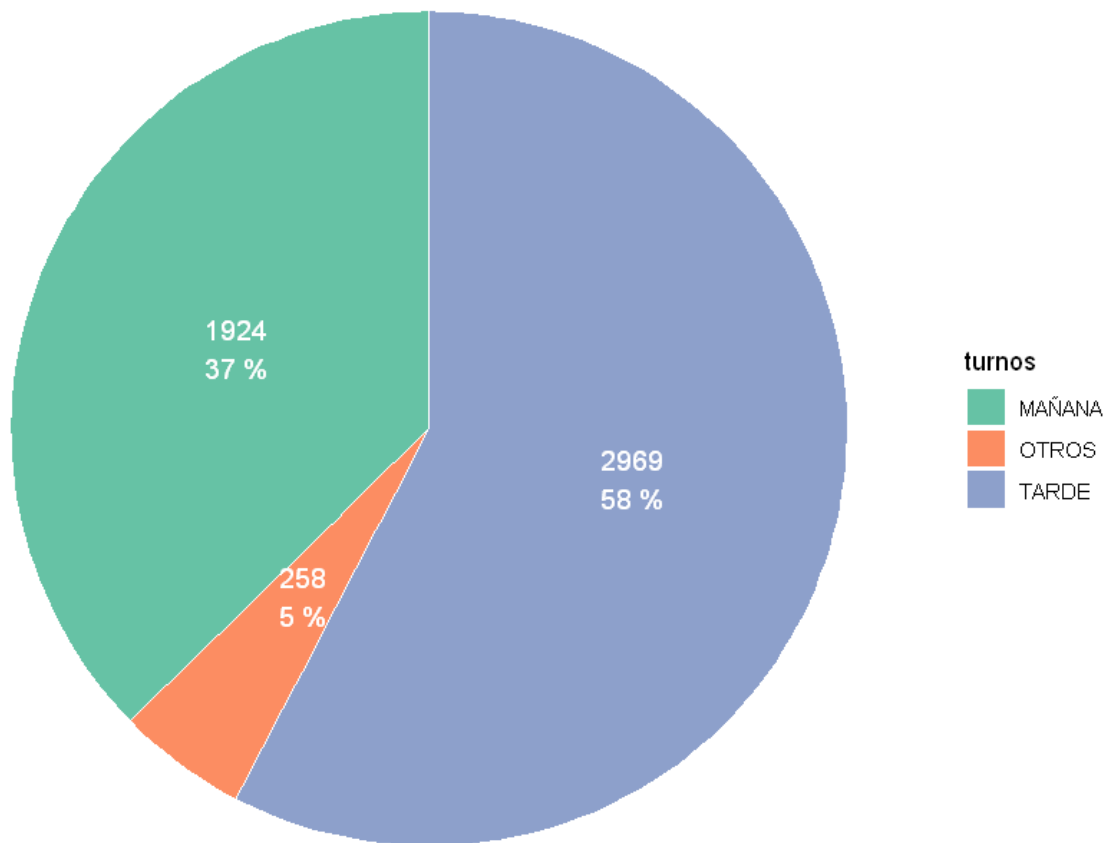


Fuente: autoría propia

Además analizamos su distribución segmentada por turnos, donde agrupamos los turnos “Intermedio” (frecuencia 10, porcentaje 0%), “Completo” (frecuencia 631, porcentaje 2%) y “Rotativo” (frecuencia 785, porcentaje 3%) en el segmento “Otros” para una mejor lectura del gráfico 2, ya que los valores porcentuales son muy pequeños y se superponen entre sí. En este podemos observar que la mayoría de estudiantes pertenecen al turno tarde con 57%, seguido del turno mañana con 38% y por último otros con 5%.

Gráfico 2

Distribucion Porcentual de la Matricula de Primaria de 2022 en el Departamento La Paz por Turnos

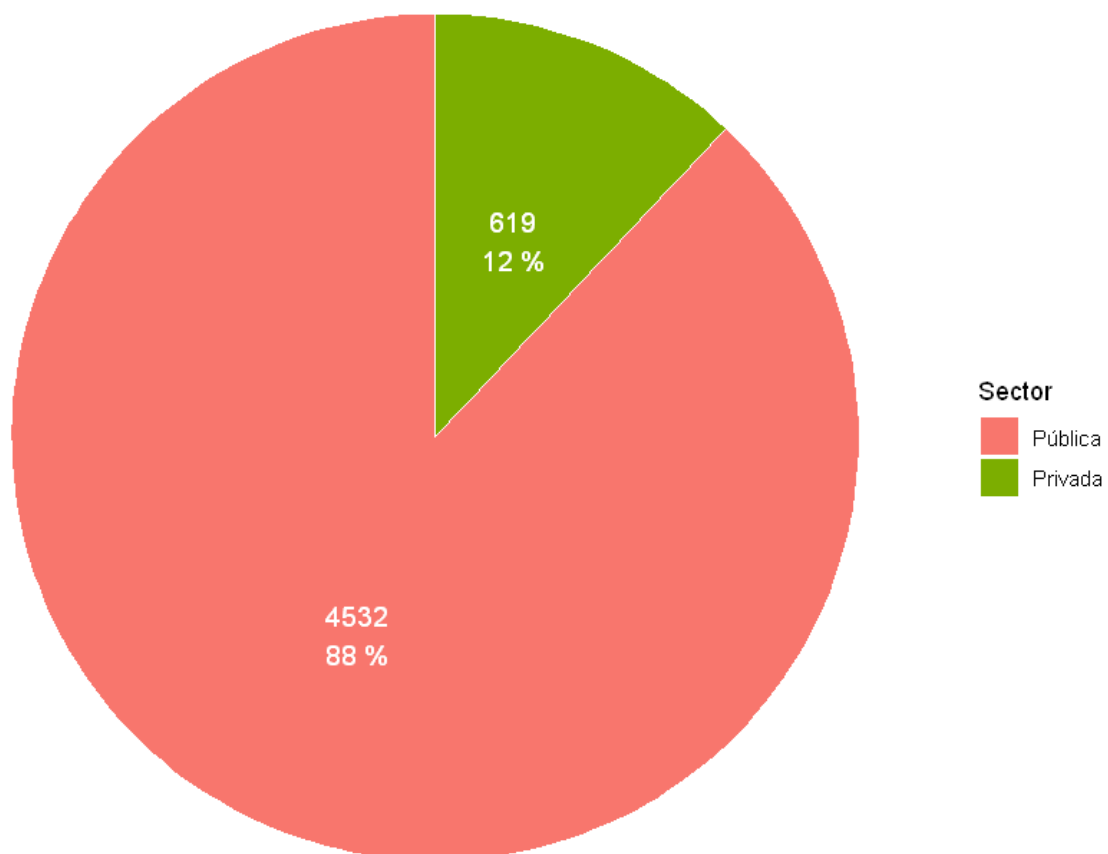


Fuente: autoría propia

Por último, analizamos la distribución segmentada por escuelas públicas y privadas de las notas cargadas, donde encontramos que el 88% pertenecen al sector público.

Gráfico 3

Distribucion Porcentual de Estudiantes de Primaria Segmentado por Sector



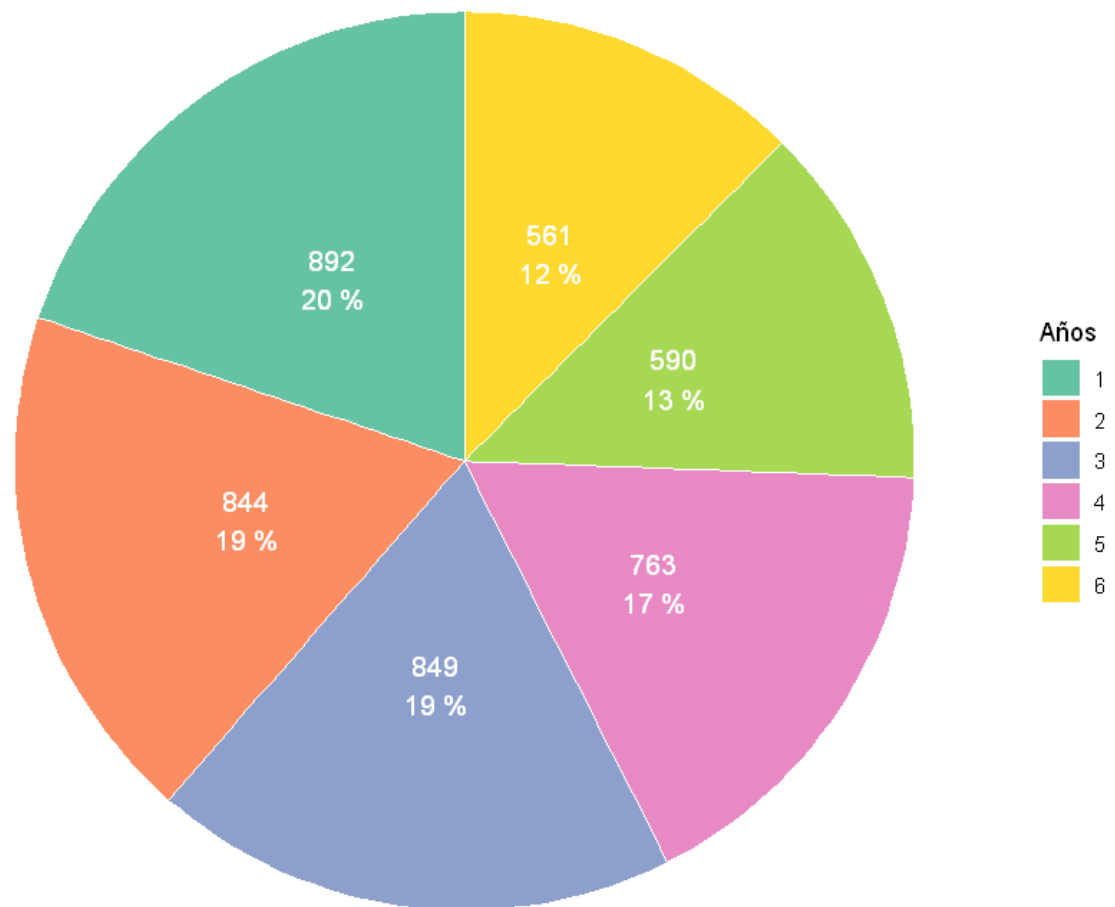
Fuente: autoría propia

Ahora procedemos con el mismo análisis para las escuelas secundarias.

Podemos ver en el gráfico 4 que el mayor porcentaje se encuentra en cuarto año con 22% seguido por quinto año con 20% , tercer y sexto año con 15 % y por último segundo y primer año con 14 %. También podemos notar la nula injerencia porcentual de los alumnos de séptimo año de escuelas técnicas debido a su muy baja cantidad de matrículas

Gráfico 4

Distribucion Porcentual de la Matricula de secundaria de 2022 en departamento La Paz por Año de Cursado

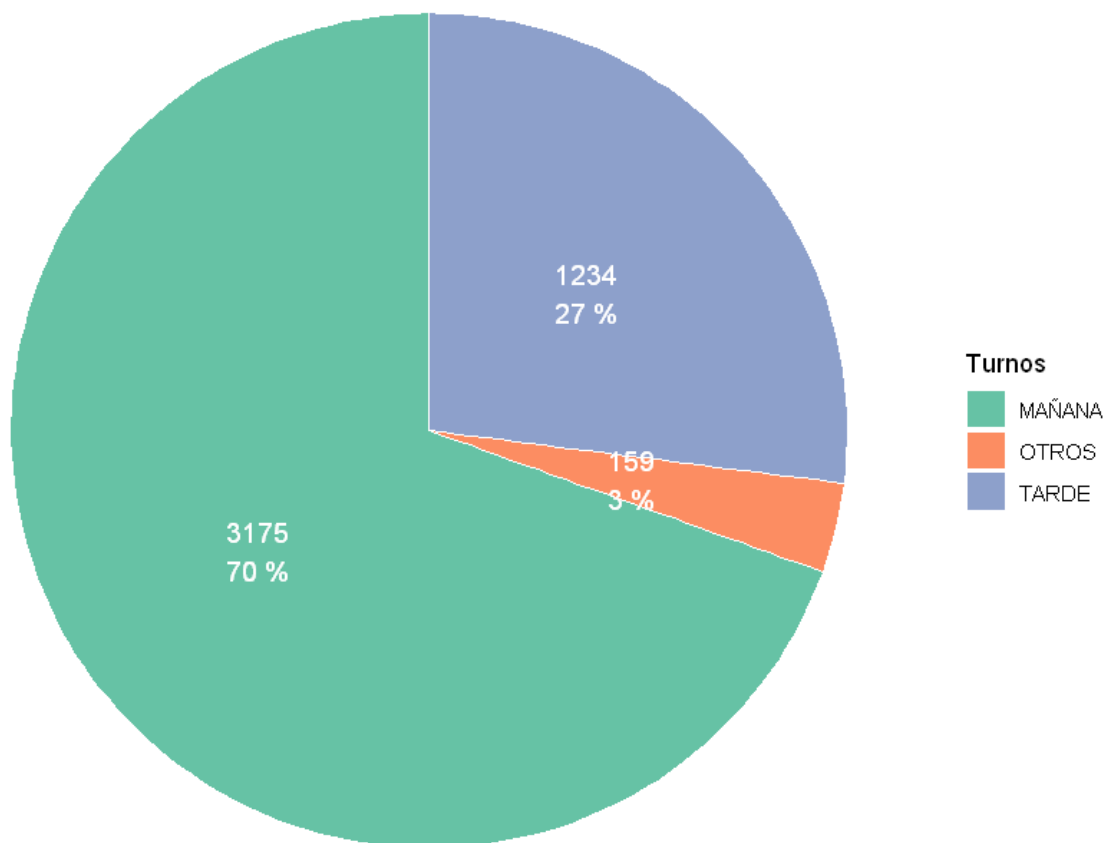


Fuente: autoría propia

Para la distribución por turnos de escuelas secundarias, para una mejor legibilidad del gráfico 5 agrupamos los turnos diurno, noche, rotativo y vespertino en la categoría otros. Podemos observar en el gráfico 5, al contrario de lo que sucedía con las escuelas primarias, el turno predominante es el turno mañana con 77 %.

Gráfico 5

Distribucion Porcentual de la Matricula de Secundaria de 2022 en el Departamento La Paz por Turnos



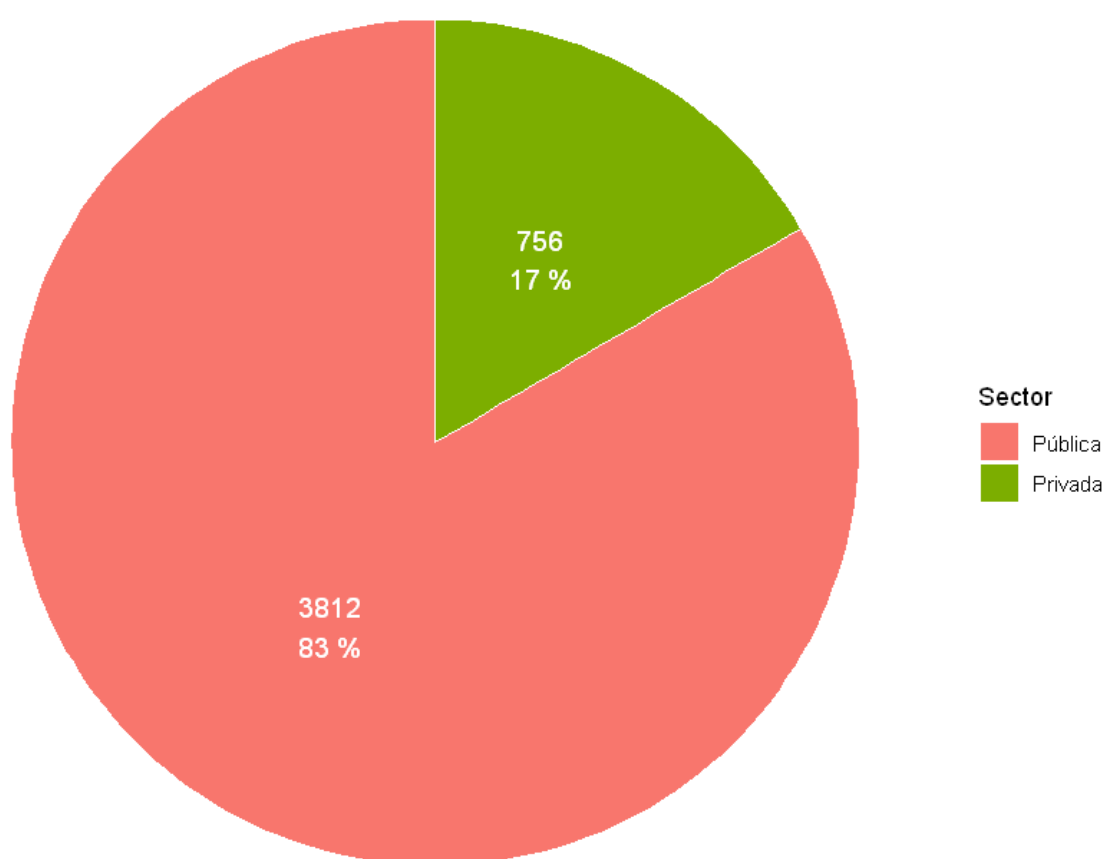
Fuente: Autoría propia

Fu

En el gráfico 6 podemos observar la distribución de las notas cargadas para escuelas secundarias segmentada por sector público y privado. Nótese que la distribución es parecida a la distribución de escuelas primarias.

Gráfico 6

Distribución Porcentual de Estudiantes de Secundaria Segmentado por Sector



Fuente: autoría propia

Rendimiento académico de matemática y lengua segmentado por año académico.

En el apartado anexo se encuentran los gráficos 7 y 8 (1°), 9 y 10 (2°), 11 y 12 (3°), 13 y 14 (4°), 15 y 16 (5°), 17 y 18 (6°) correspondientes a los histogramas segmentados por año académico para el promedio de matemática y lengua de las escuelas primarias.

Donde podemos notar que.....

Pasando a las escuelas secundarias en el apartado anexo se encuentran los gráficos 19 y 20 (1 año), 21 y 22 (2 año), 23 y 24 (3 año), 25 y 26 (4 año), 27 y 28 (5 año), 29 y 30 (6 año) correspondientes a los histogramas segmentados por año académico para el promedio de matemática y lengua y literatura de las escuelas secundarias.

Podemos observar que el caso de secundaria es más variado que primaria....

Primaria: matemática

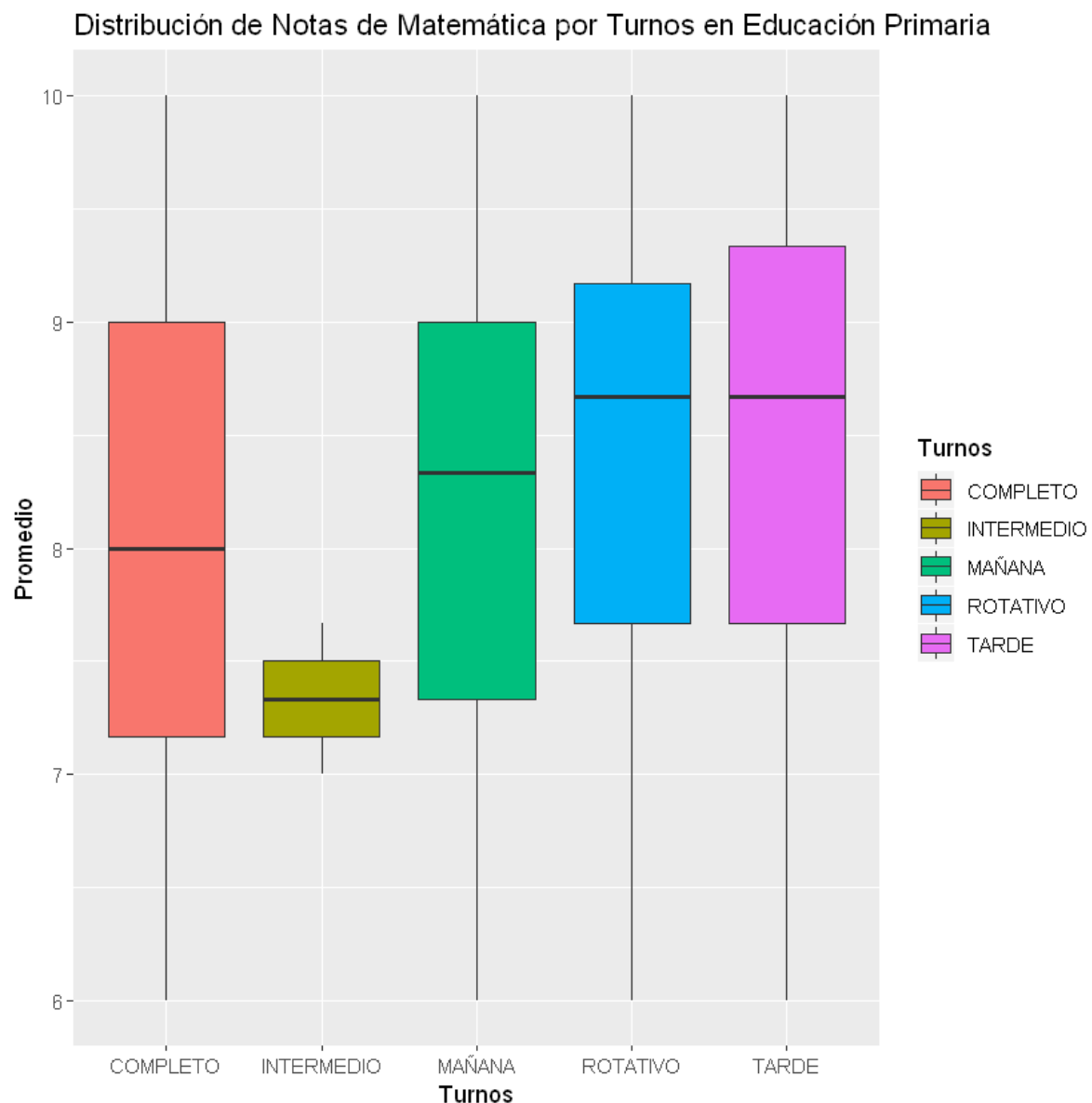
Tabla 3

Tabla de frecuencia de promedios segmentado en turnos para la materia de matemáticas

COMPLETO	INTERMEDIO	MAÑANA	ROTATIVO	TARDE
118	2	1936	157	2999

En el gráfico 31 podemos observar que el rango, exceptuando el turno intermedio, se encuentra entre 6 y 10. También se observa que la mediana se encuentra entre [8,9) para todos los turnos exceptuando intermedio.

Gráfico 31



Primaria: Lengua

Tabla 4

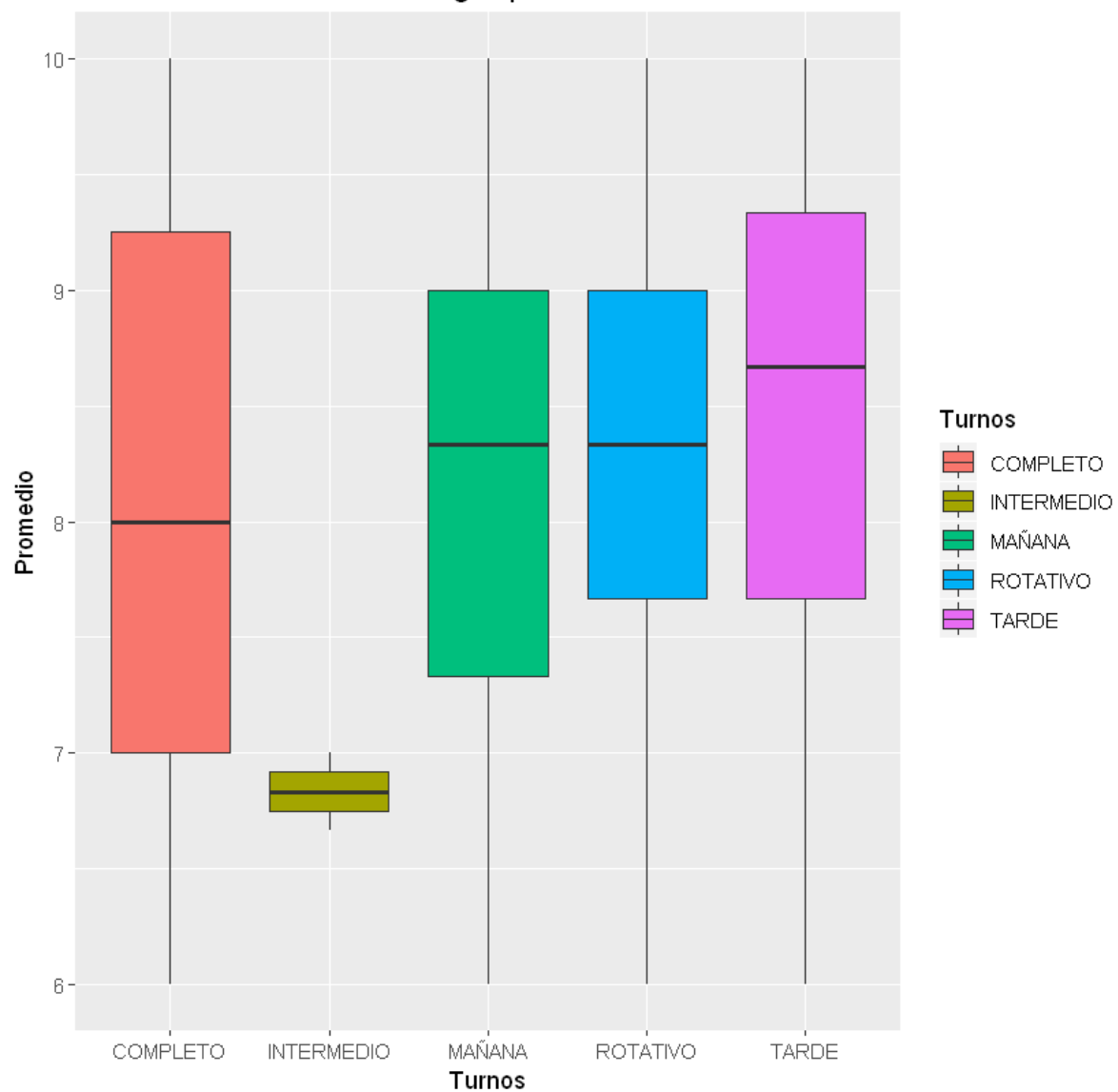
Tabla de frecuencia de promedios segmentado en turnos para la materia de lengua

COMPLETO	INTERMEDIO	MAÑANA	ROTATIVO	TARDE
118	2	1964	158	2990

En el gráfico 32 vemos que, al igual que sucedía con los promedios de matemática, la mediana para todos los turnos se encuentra entre [8,9), exceptuando el turno intermedio.

Gráfico 32

Distribución de Notas de Lengua por Turnos en Educación Primaria



Podemos observar en la tabla 3 como en la tabla 4, que la cantidad de notas cargadas para los turnos “Completo”, “Intermedio” y “Rotativo” no es significativa a comparación del turno

“Mañana” y “Tarde”, por lo que la información de ellos en estos gráficos puede ser imprecisa.

Secundaria: matemática

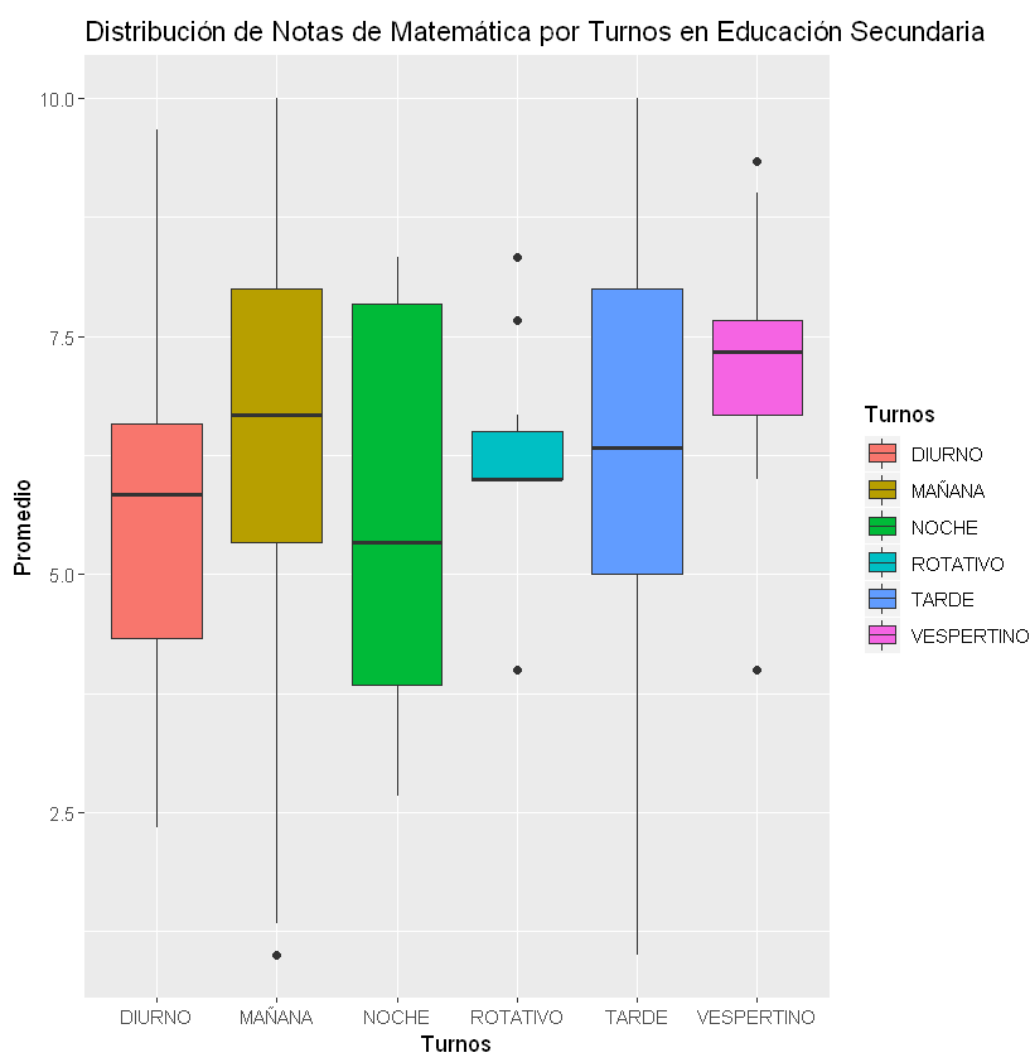
Tabla 5

Tabla de frecuencia de promedios segmentado en turnos para la materia de matemática

DIURNO	MAÑANA	NOCHE	ROTATIVO	TARDE	VESPERTINO
91	5727	16	11	1259	21

El gráfico 33 muestra la distribución de los promedios de matemática segmentado por turnos donde podemos notar que el turno mañana y tarde hay más variabilidad. En el turno mañana el 50% tiene un promedio de 6,5 aproximadamente, mientras que el turno de la tarde tiene un promedio de 6 aproximadamente.

Gráfico 33



Secundaria: Lengua y literatura

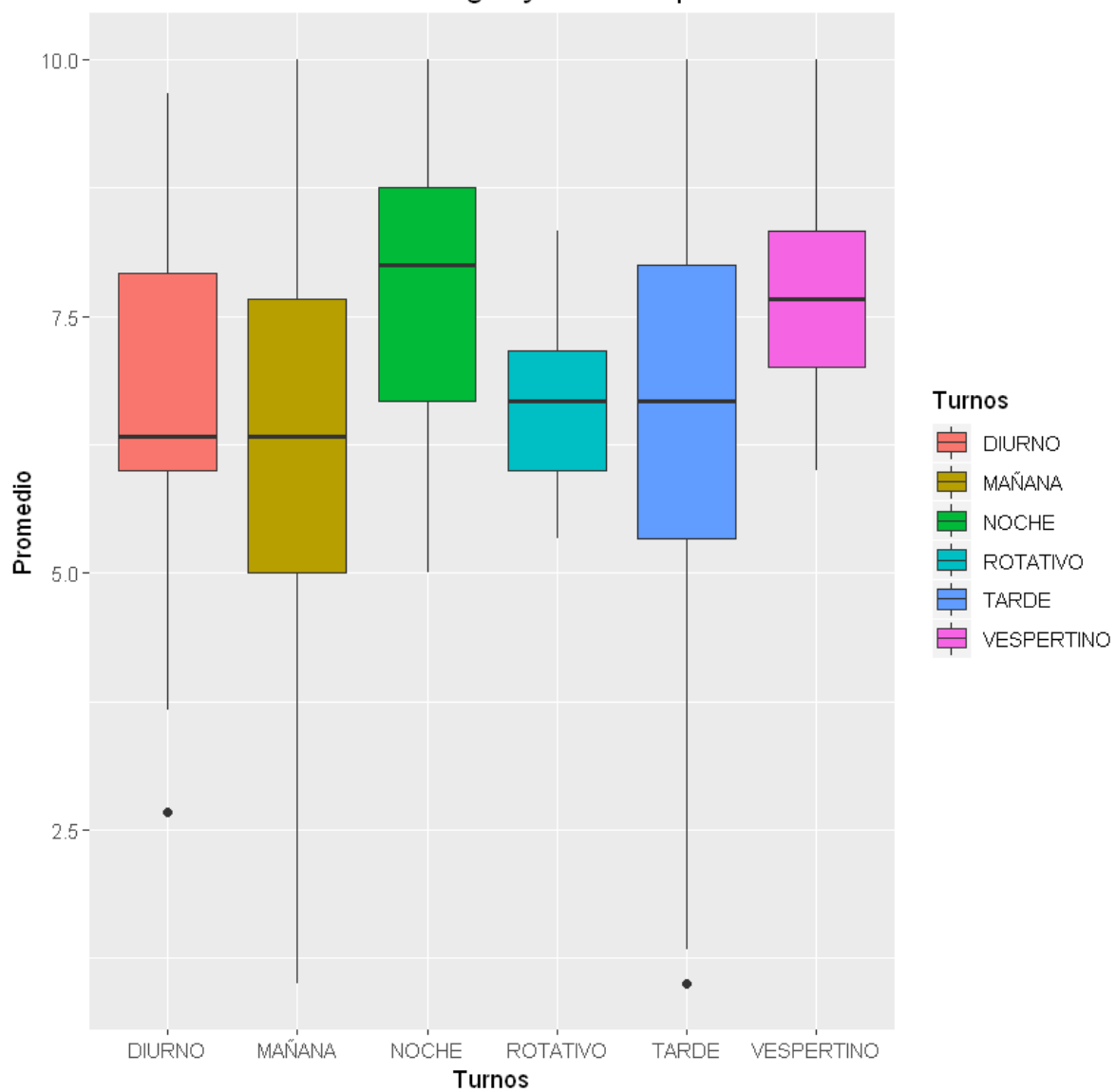
Tabla 6

Tabla de frecuencia de promedios segmentado en turnos para la materia de lengua y literatura

DIURNO	MAÑANA	NOCHE	ROTATIVO	TARDE	VESPERTINO
86	2829	52	11	1055	21

Gráfico 34

Distribución de Notas de Lengua y Literatura por Turnos en Secundaria



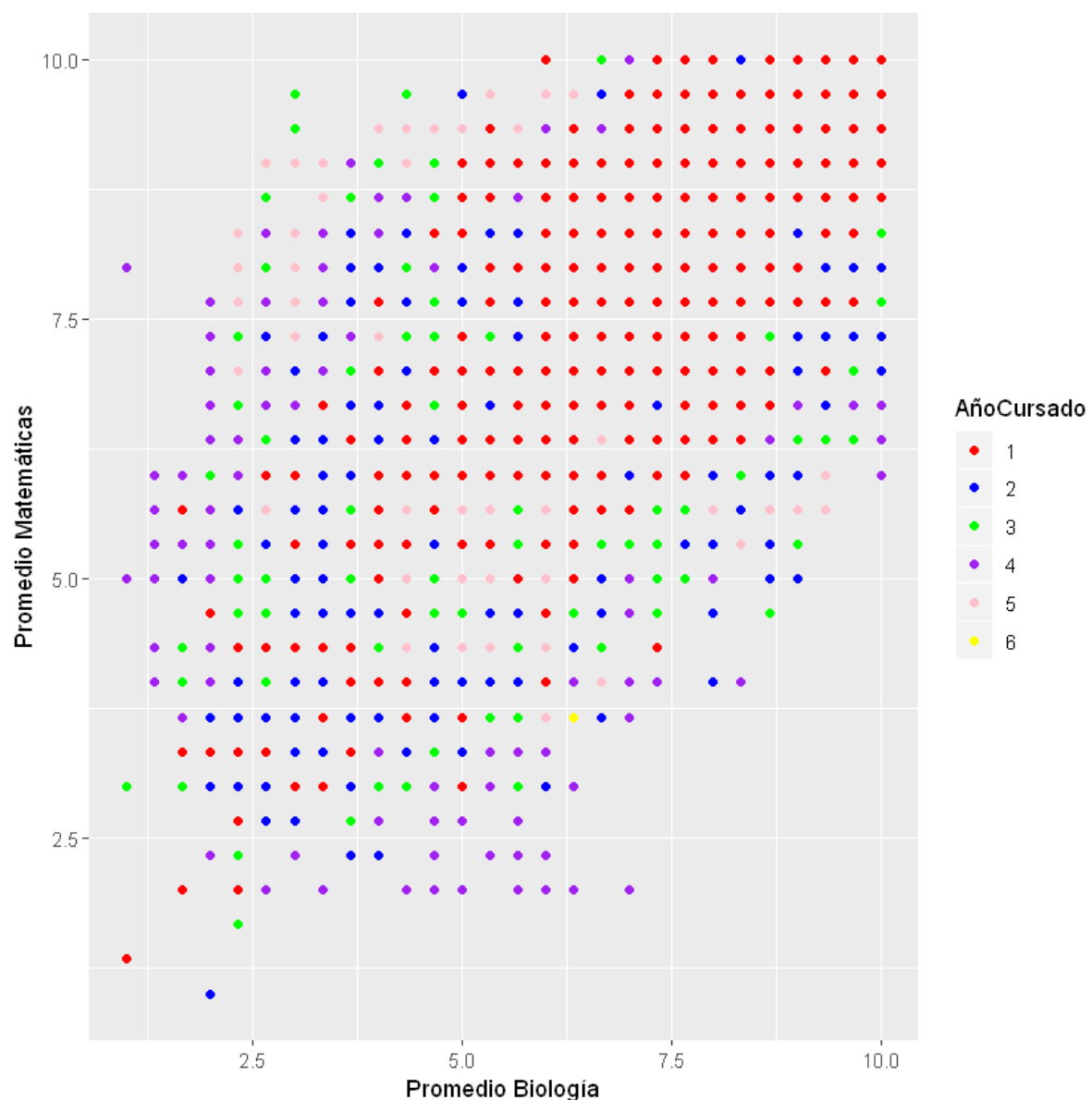
Podemos observar en la tabla 5 como en la tabla 6, que la cantidad de notas cargadas para los turnos “Diurno”, “Noche”, “Rotativo” y “Vespertino” no es significativa a comparación del

turno “Mañana” y “Tarde”, por lo que la información de los gráficos para estos turnos puede ser imprecisa.

Correlaciones

Se plantea como hipótesis que si el rendimiento de un alumno de un mismo año en la materia de Matemáticas es bueno, será de igual manera para la materia de Biología. Por consiguiente, realizamos un diagrama de dispersión entre los promedios de estas materias diferenciado por año para ver el comportamiento entre las dos variables, donde tomaremos solamente las escuelas secundarias ya que es el dataset que presenta más variación en los datos.

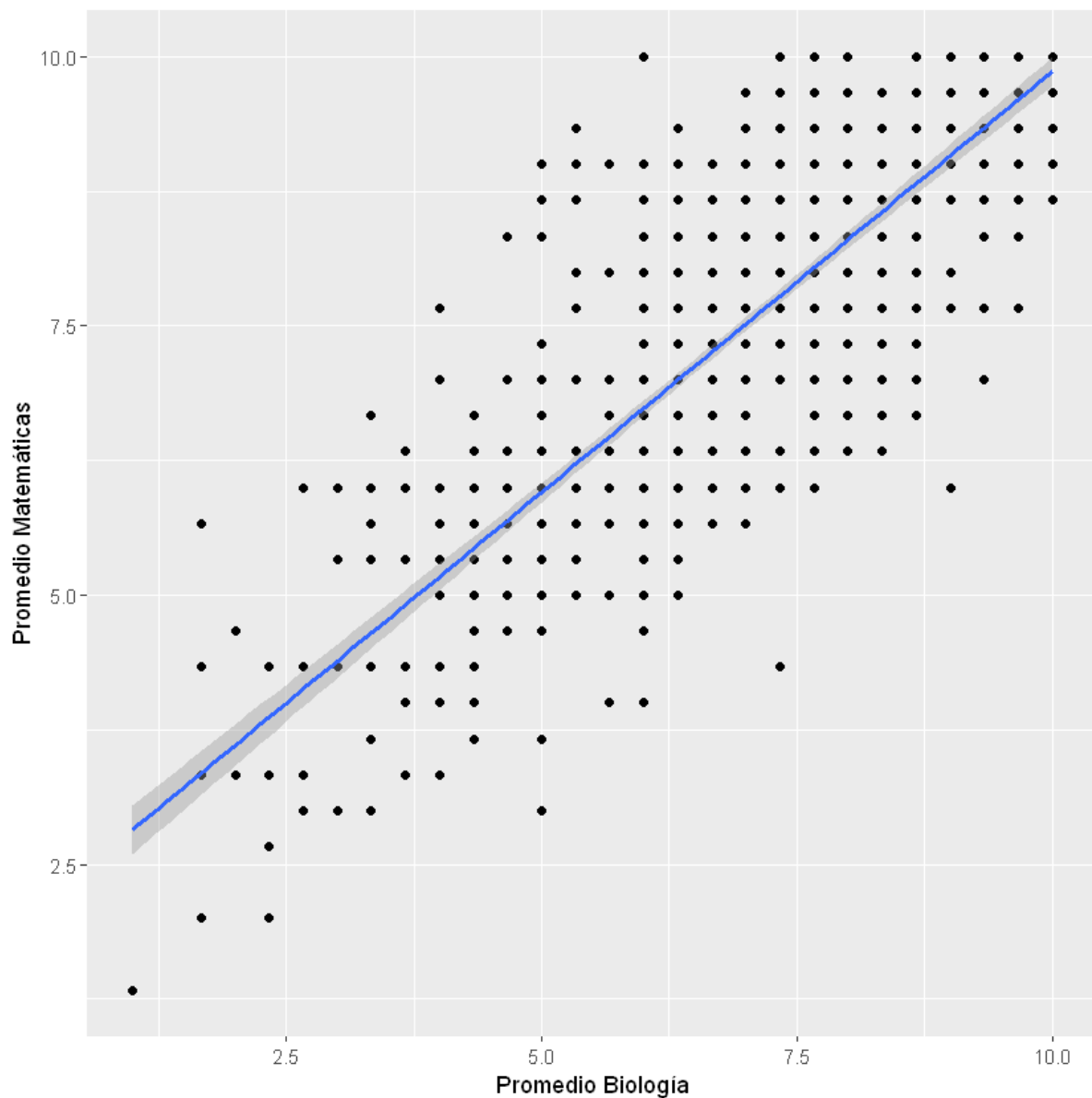
Gráfico 35



Como podemos observar en el Gráfico 35, el primer año de cursado es el que aparentemente presenta una relación lineal entre ambas variables, por lo que a partir de

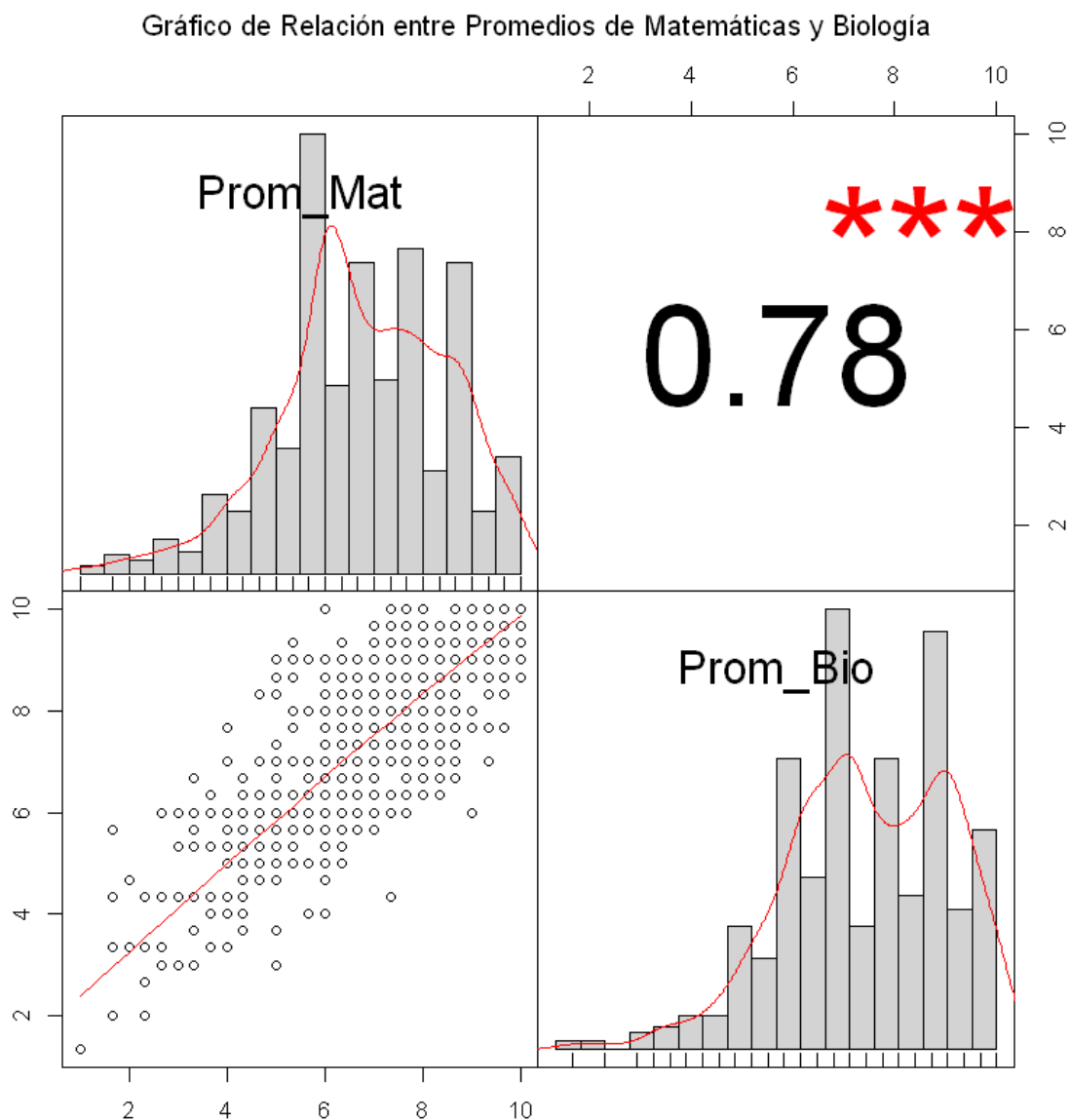
este momento, centraremos el estudio solamente en este año. Volvemos a graficar, pero esta vez solo con los datos del primer año (Gráfico 36).

Gráfico 36



Para verificar que efectivamente las variables estén correlacionadas, debemos realizar un test de normalidad para saber qué tipo de test utilizar para verificar la correlación, y como nuestra muestra tiene más de 50 elementos, utilizaremos el test de Lilliefors (Kolmogorov-Smirnov) para normalidad para cada una de las variables con una confianza del 95%. Al hacerlo, el test para ambas variables da como resultado un p-value menor a 0.05, por lo que podemos afirmar que las variables no siguen una distribución normal. Como consecuencia, utilizaremos el método de Spearman para comprobar la correlación entre las variables, y, para ello, lo haremos de forma gráfica.

Gráfico 37



Observando el Gráfico 37, el nivel de significancia es alto (***) y el ρ de Spearman da 0.78, por lo que se puede afirmar que tenemos una buena relación.

Ahora bien, procederemos a agregar una nueva variable a nuestro estudio, siendo esta Lengua y Literatura, para comprobar que la correlación encontrada no sea espuria. Para ello hallamos la matriz de correlaciones parciales utilizando los promedios de Matemáticas, Biología y Lengua y Literatura.

El resultado de esto puede verse en el Gráfico 38, donde podemos ver que al agregar Lengua y Literatura, la buena correlación que existía entre Matemáticas y Biología ahora es baja, por lo que podemos decir que la correlación estaba enmascarada con la relación de Lengua y Literatura. Queda por ver si la correlación entre Lengua y Biología, o entre Lengua y Matemáticas son significativas analizandolas por sí mismas.

Gráfico 38

\$estimate

	Prom_Mat	Prom_Bio	Prom_Len
Prom_Mat	1.0000000	0.3266264	0.4784281
Prom_Bio	0.3266264	1.0000000	0.5280509
Prom_Len	0.4784281	0.5280509	1.0000000

\$p.value

	Prom_Mat	Prom_Bio	Prom_Len
Prom_Mat	0.000000e+00	6.389809e-30	1.159734e-66
Prom_Bio	6.389809e-30	0.000000e+00	2.358793e-83
Prom_Len	1.159734e-66	2.358793e-83	0.000000e+00

\$statistic

	Prom_Mat	Prom_Bio	Prom_Len
Prom_Mat	0.00000	11.69369	18.43581
Prom_Bio	11.69369	0.00000	21.04079
Prom_Len	18.43581	21.04079	0.00000

\$n

1148

\$gp

1

\$method

'spearman'

Conclusiones

En primer lugar y dado al análisis exploratorio realizado estamos en condiciones de decir que en la implementación del boletín virtual hace falta una mejora a la hora de cargar datos ya que en un principio el estado general de la base de datos era malo y en consecuencia demanda tiempo en la limpieza y reorganización de los datos. El cual se podría destinar a un análisis más profundo.

En cuanto al análisis del rendimiento académico, en el caso de las escuelas primarias observamos en la tabla 7 que el promedio para las materias de matemática y lengua en los seis grados es aproximadamente 8, y sus mínimos en todos los casos es 6 por lo que podemos decir que el rendimiento académico para estas materias de los estudiante de primaria del departamento La Paz es muy bueno.

Tabla 7

Media de las asignaturas matemáticas y lengua para escuelas primarias.

	Matemática	Lengua
1°	8,47	8,36
2°	8,49	8,50
3°	8,56	8,53
4°	8,31	8,27
5°	8,16	8,16
6°	8,12	8,09

En un análisis similar para las escuelas secundarias, en la tabla 8 podemos observar que la media de los promedios ronda alrededor de 6 en ambas materias, distinguiéndose en lengua y literatura quinto año con una media de 7,51 y sexto año con una media 8,30. También notamos que los mínimos llegan a la nota mínima (1). Dado esto concluimos que el rendimiento académico en estas materias es regular.

Tabla 8*Media de las asignaturas matemática y lengua y literatura*

	Matemática	Lengua y Literatura
1°	6,80	6,95
2°	6,32	6,34
3°	6,21	6,51
4°	6,14	6,04
5°	6,63	7,52
6°	6,91	8,30

Anexo

Gráfico 7

Primaria - Histograma de promedio de Matemática - 1 año.

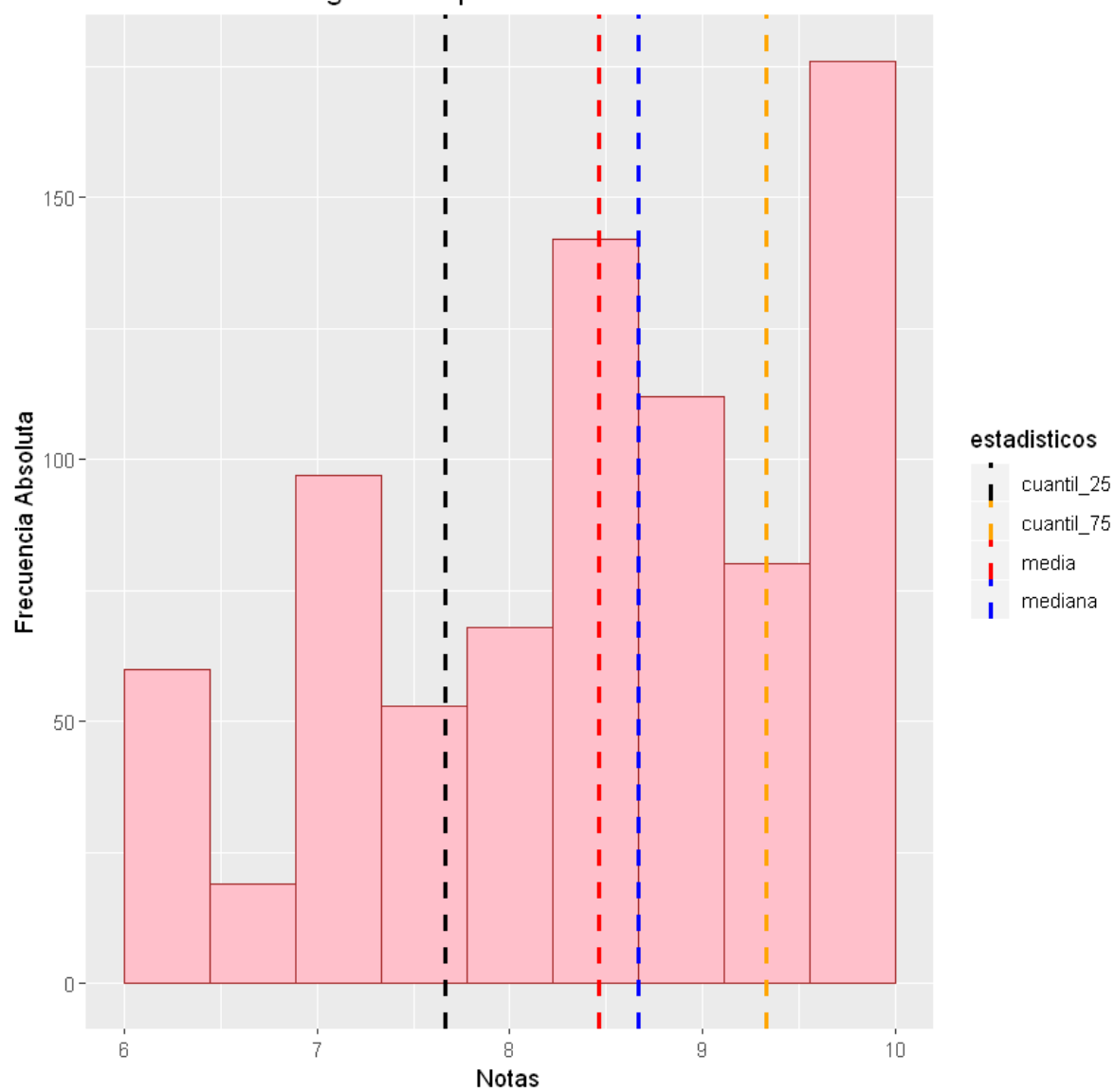


Gráfico 8

Primaria - Histograma de promedio de Lengua - 1 año.

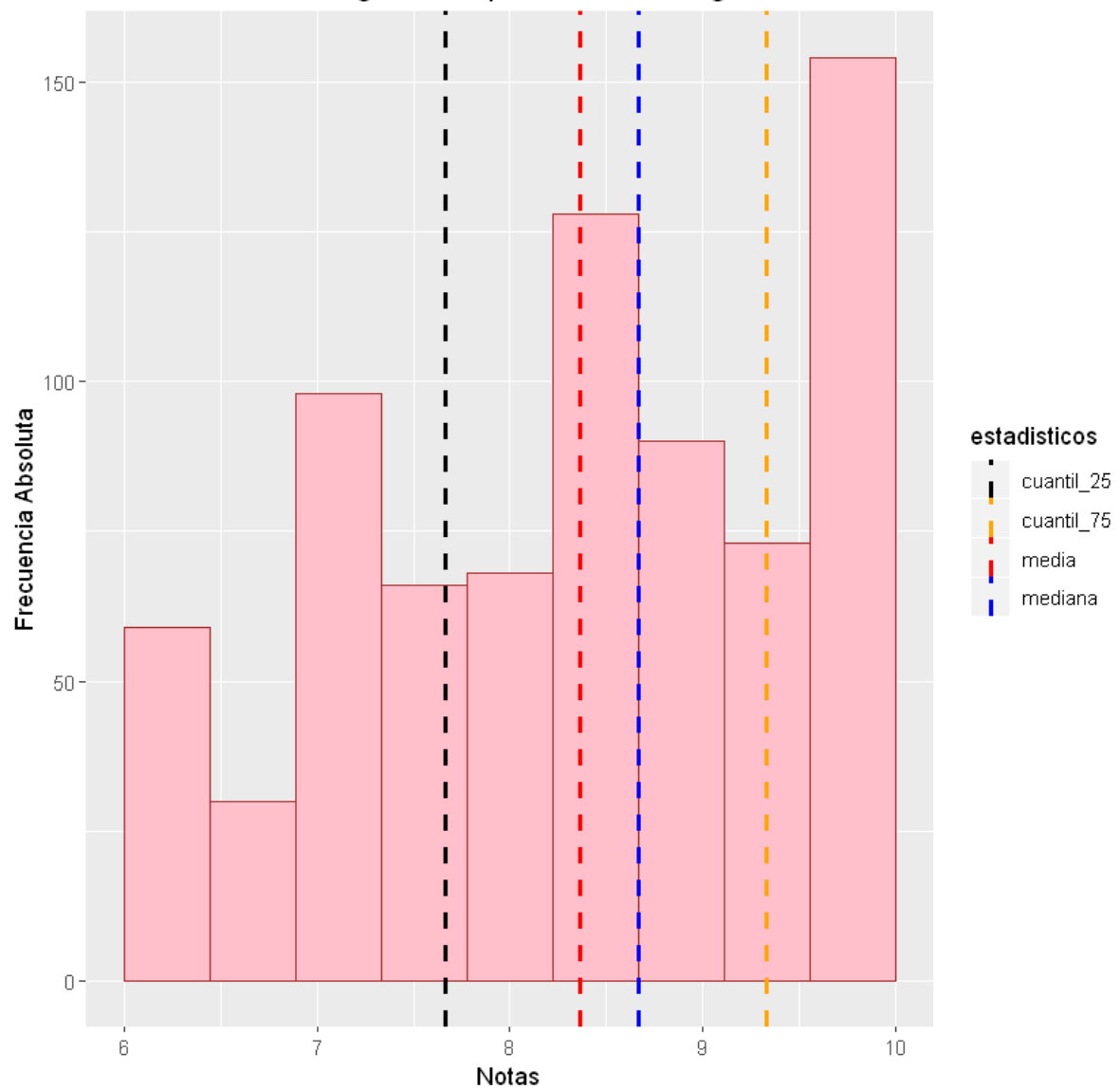


Gráfico 9

Primaria - Histograma de promedio de Matemática - 2 año.

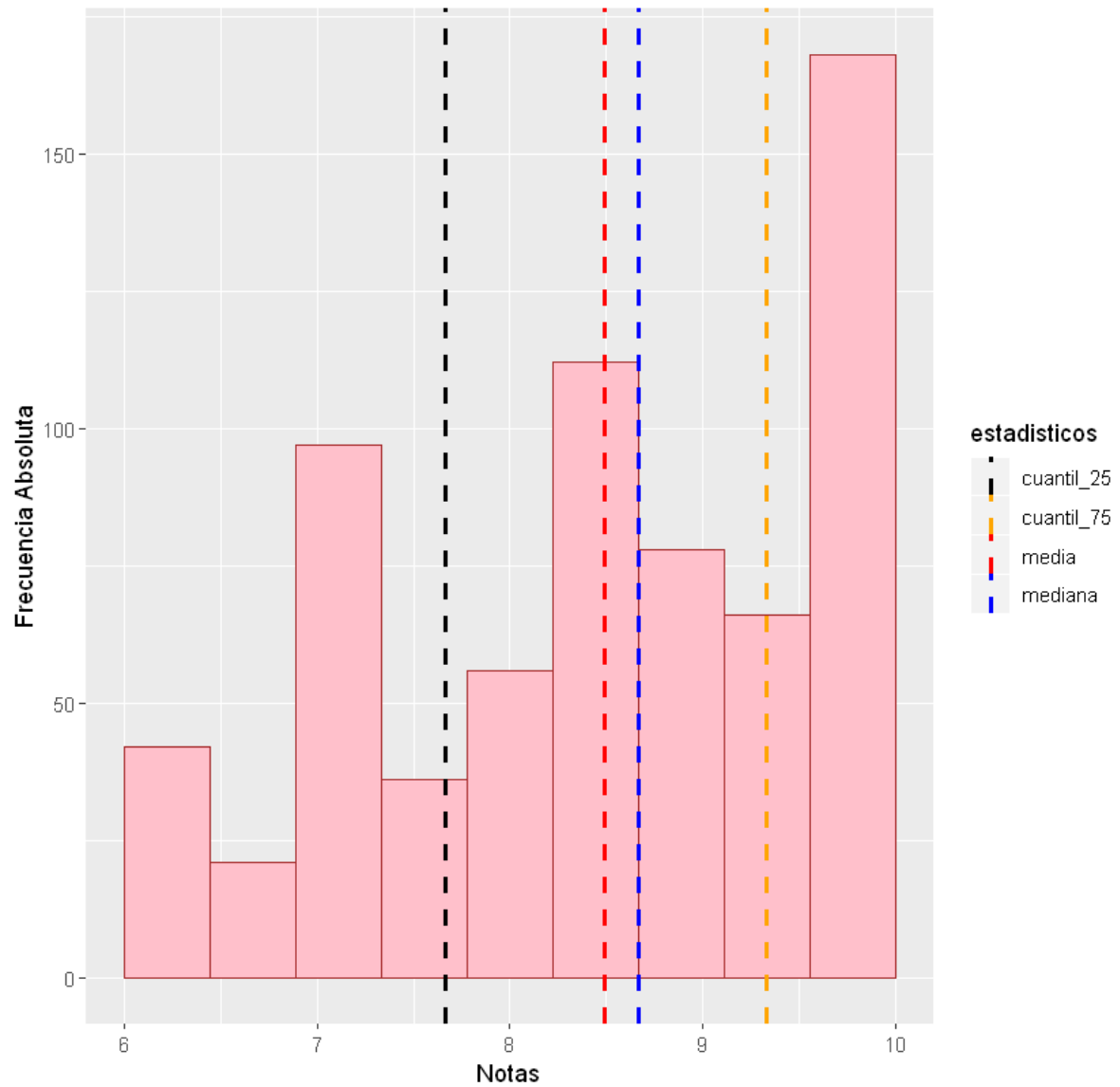


Gráfico 10

Primaria - Histograma de promedio de Lengua - 2 año.

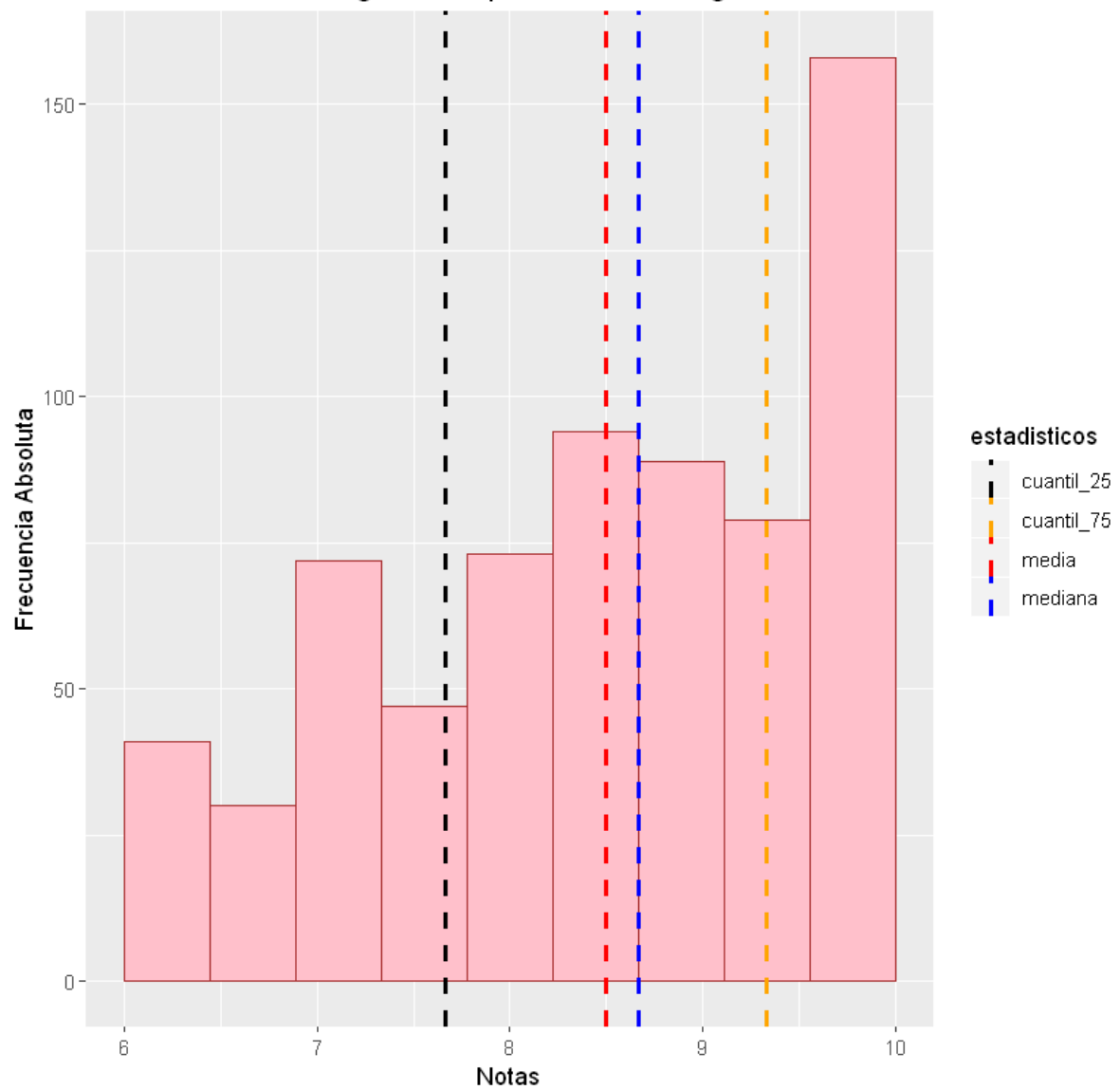


Gráfico 11

Primaria - Histograma de promedio de Matemática - 3 año.

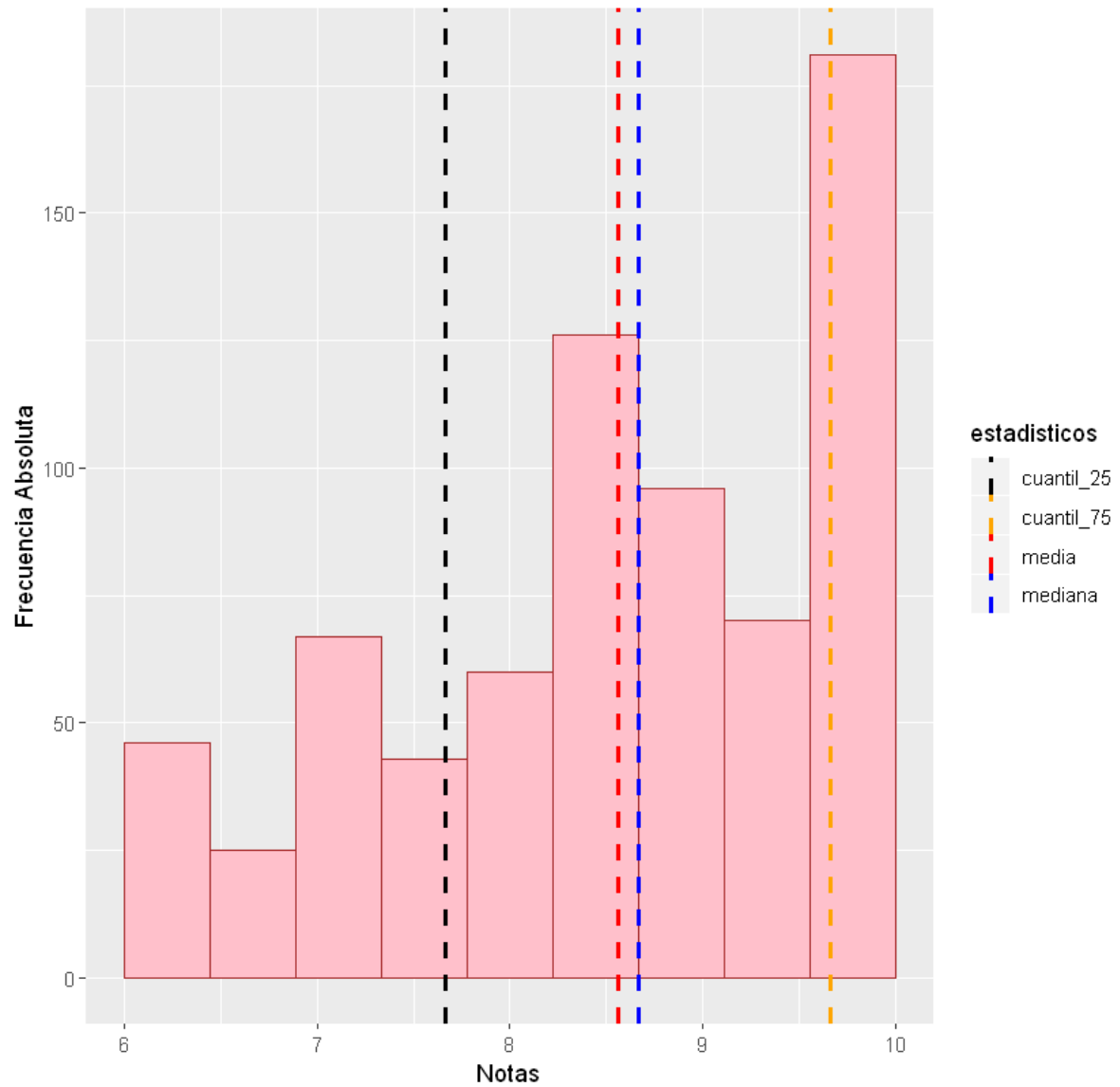


Gráfico 12

Primaria - Histograma de promedio de Lengua - 3 año.

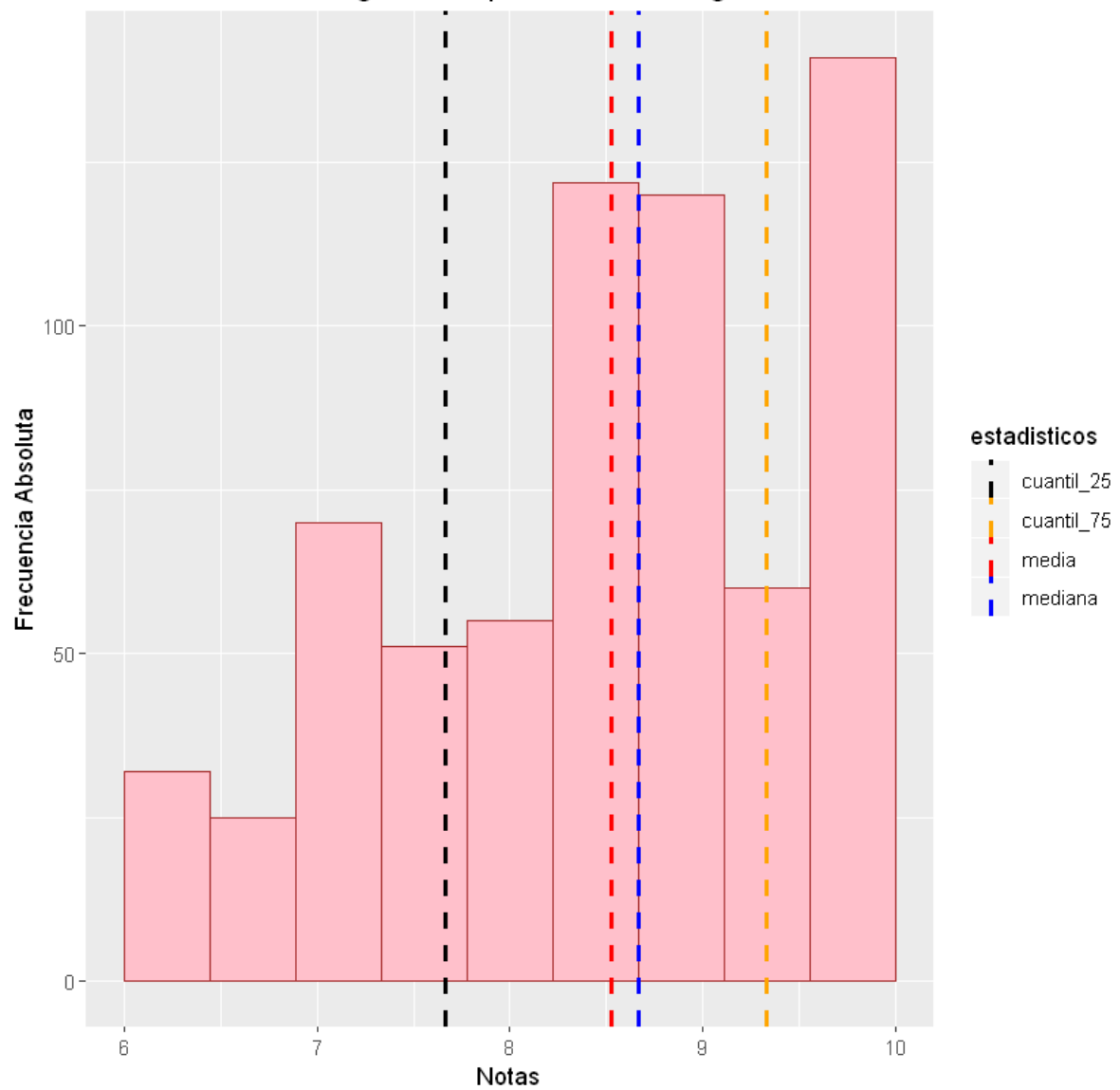


Gráfico 13

Primaria - Histograma de promedio de Matemática - 4 año.

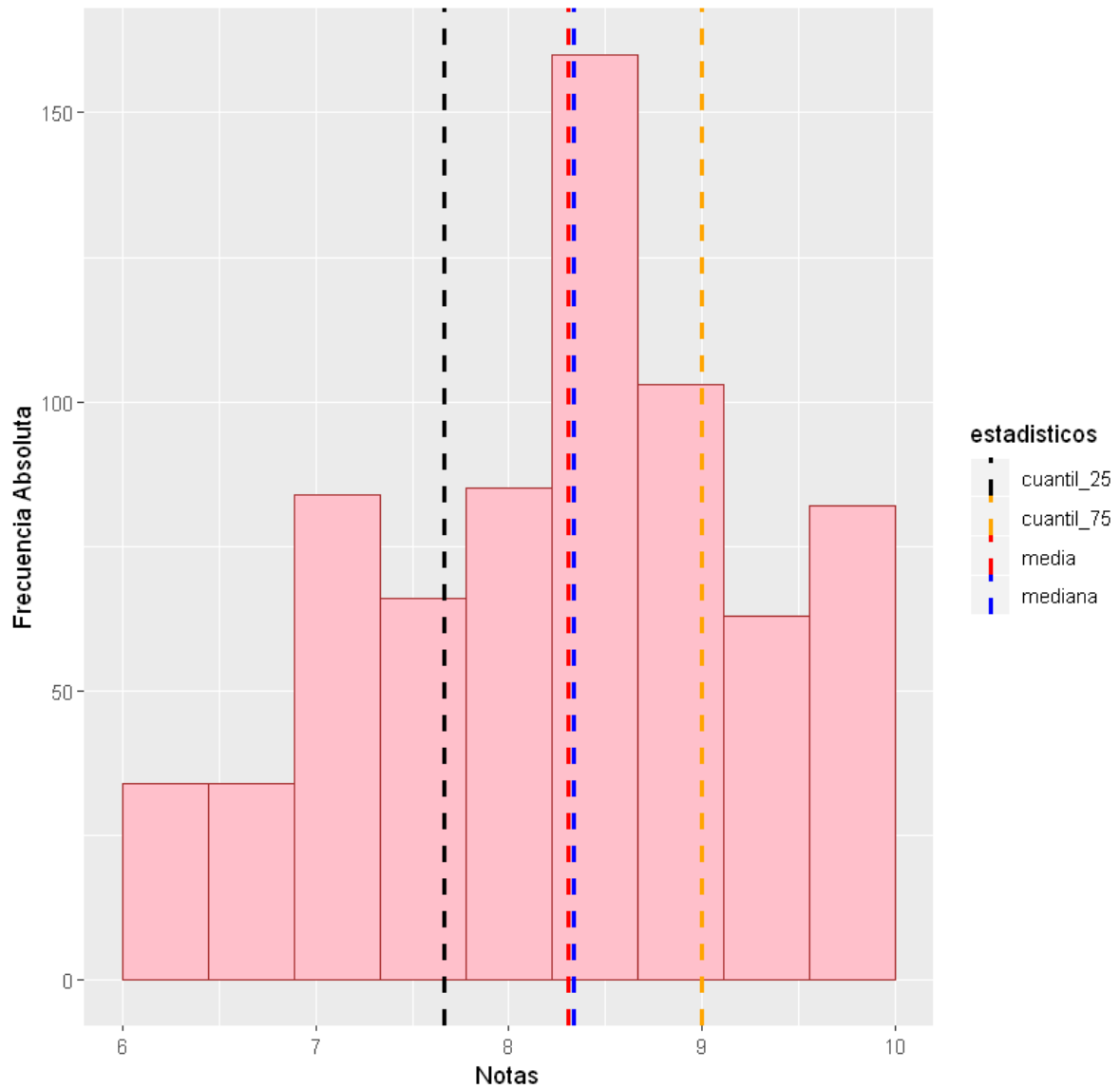


Gráfico 14

Primaria - Histograma de promedio de Lengua - 4 año.

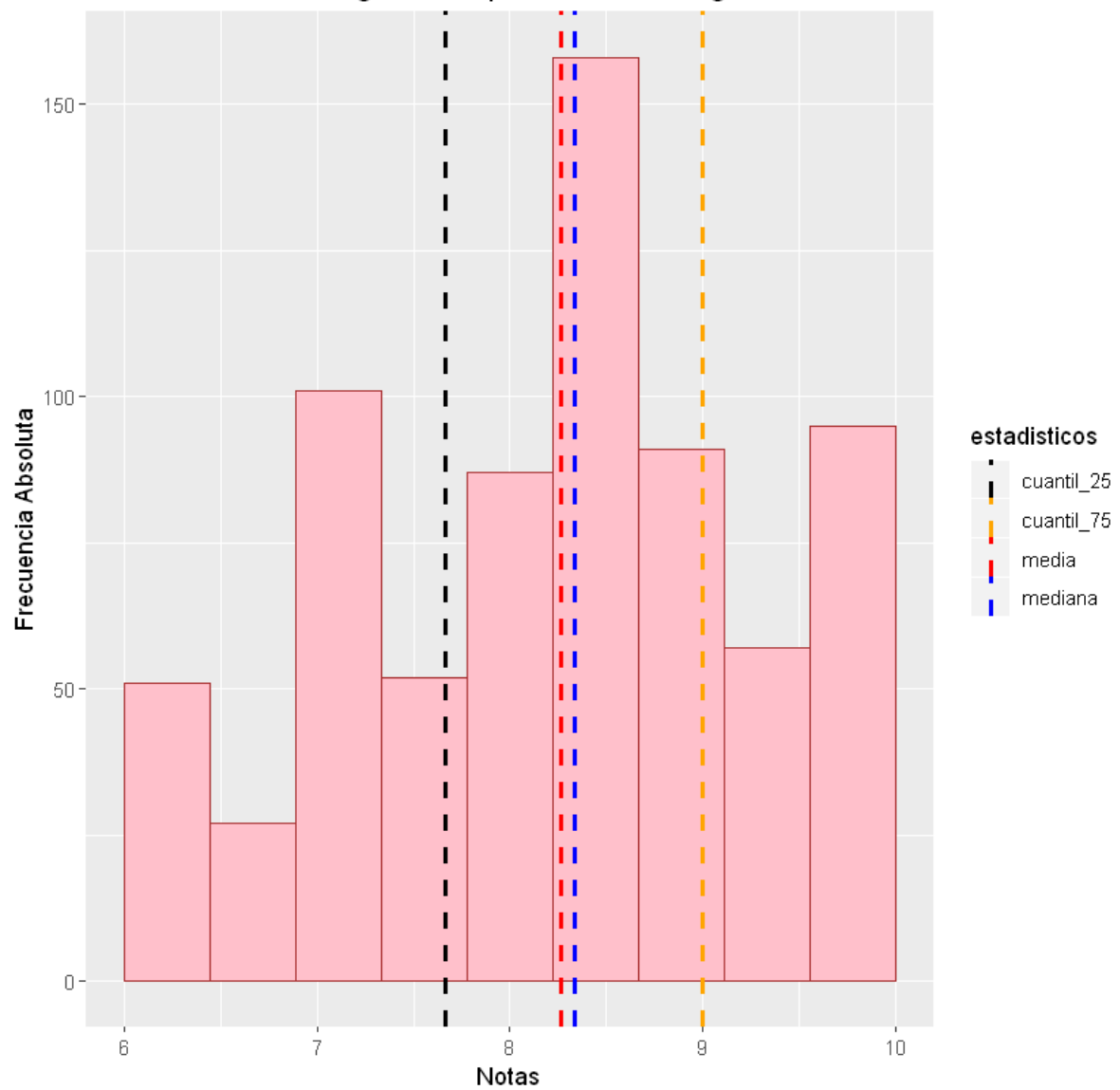


Gráfico 15

Primaria - Histograma de promedio de Matemática - 5 año.

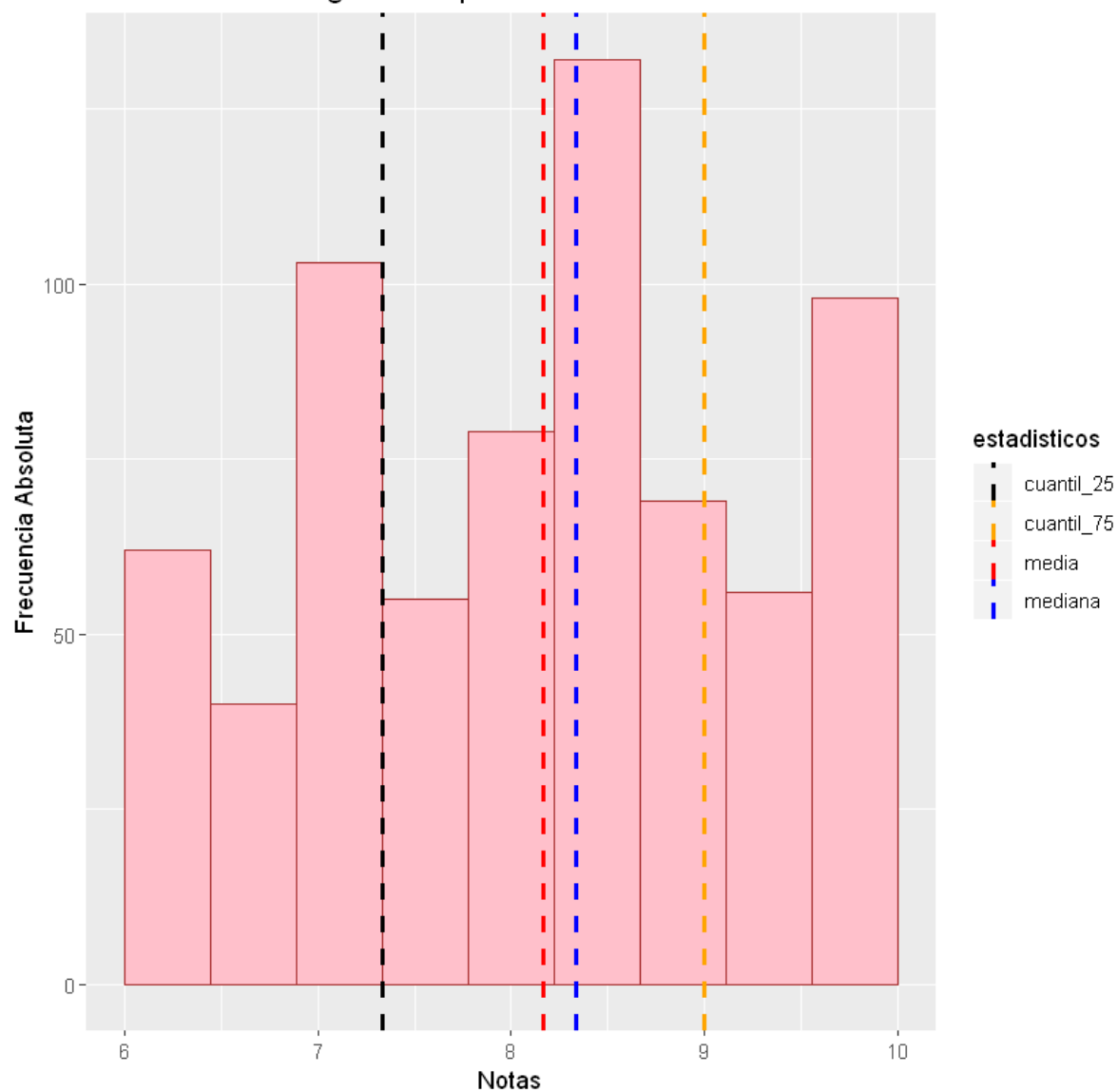


Gráfico 16

Primaria - Histograma de promedio de Lengua - 5 año.

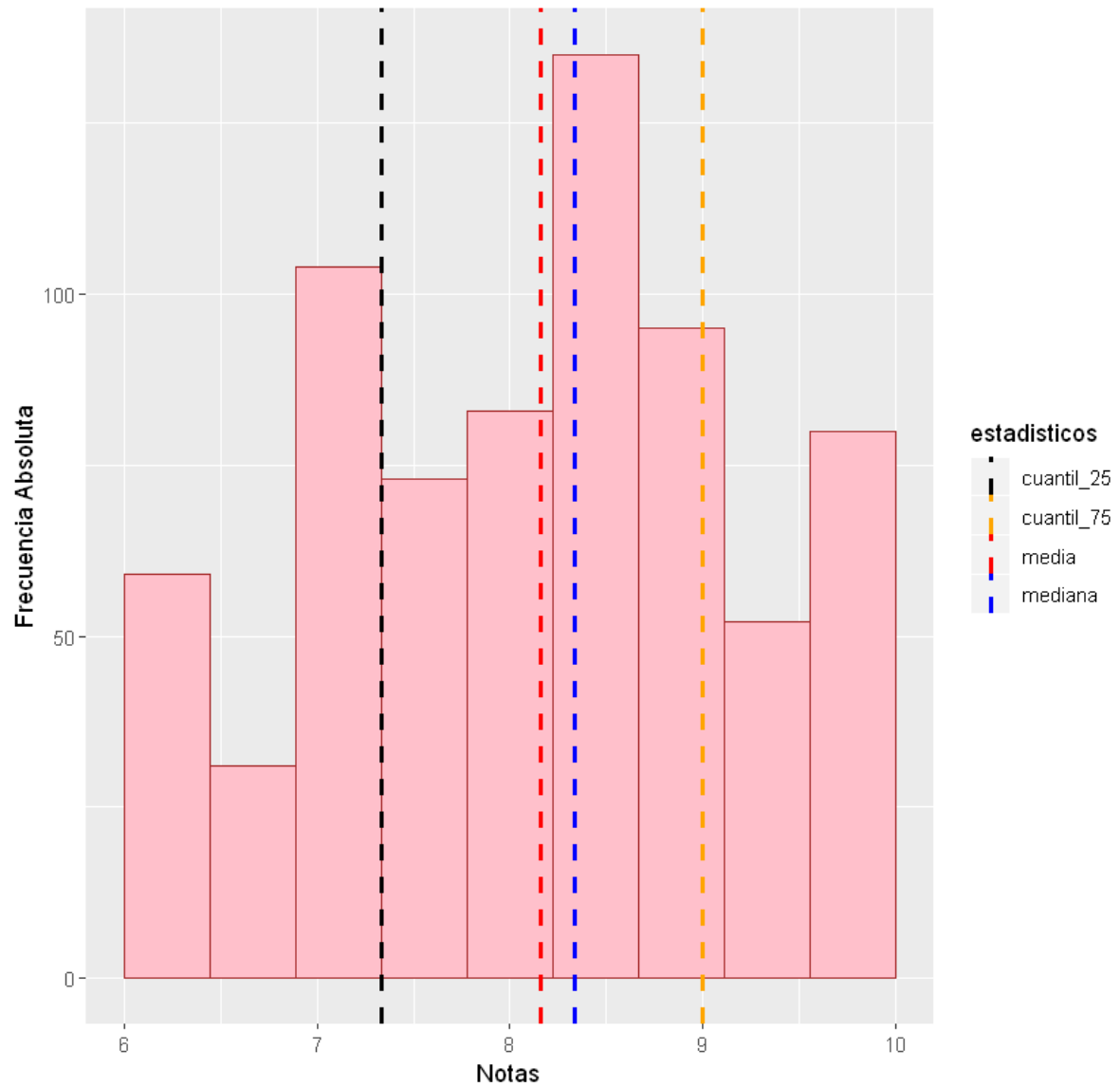


Gráfico 17

Primaria - Histograma de promedio de Matemática - 6 año.

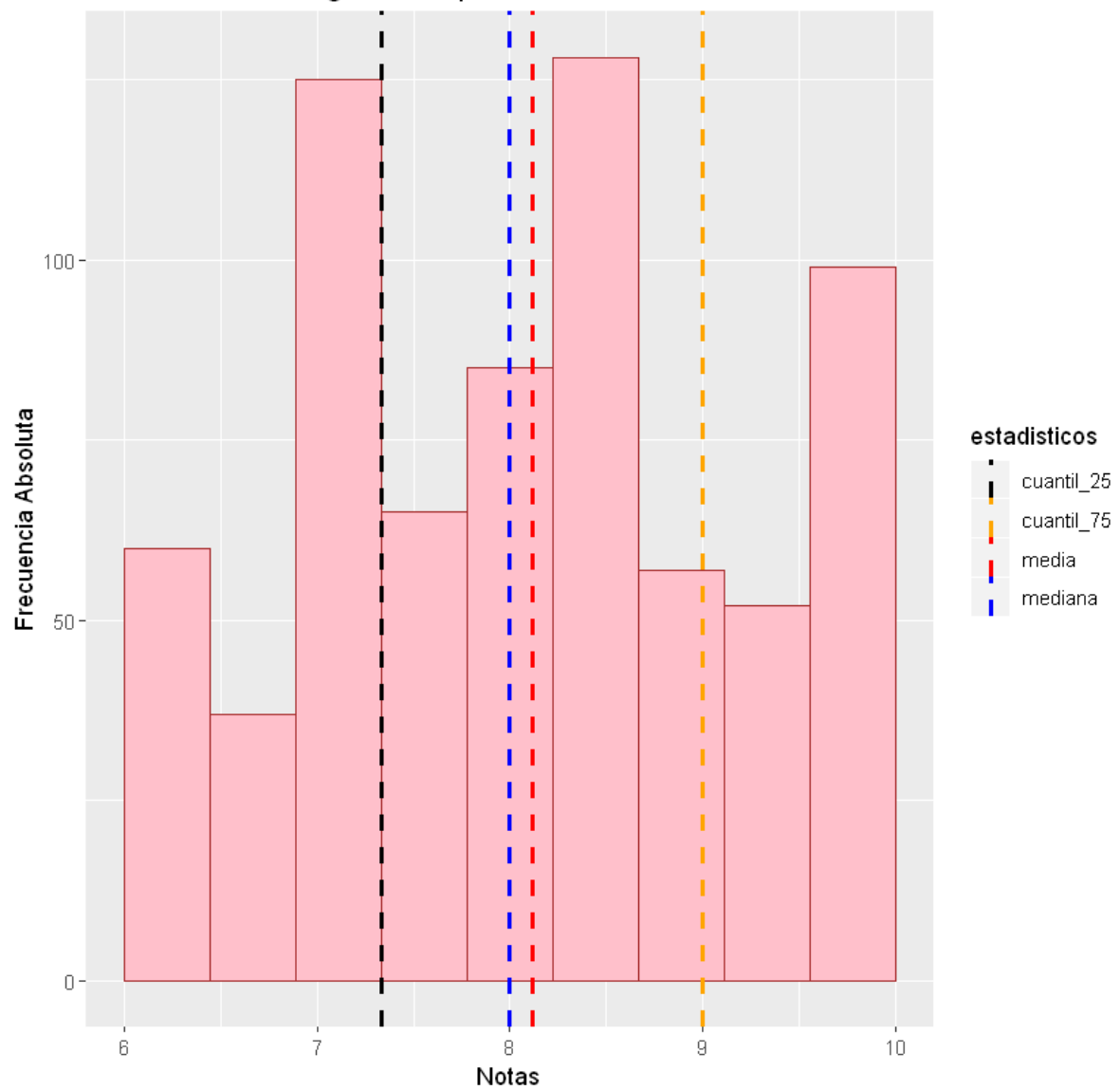


Gráfico 18

Primaria - Histograma de promedio de Lengua - 6 año.

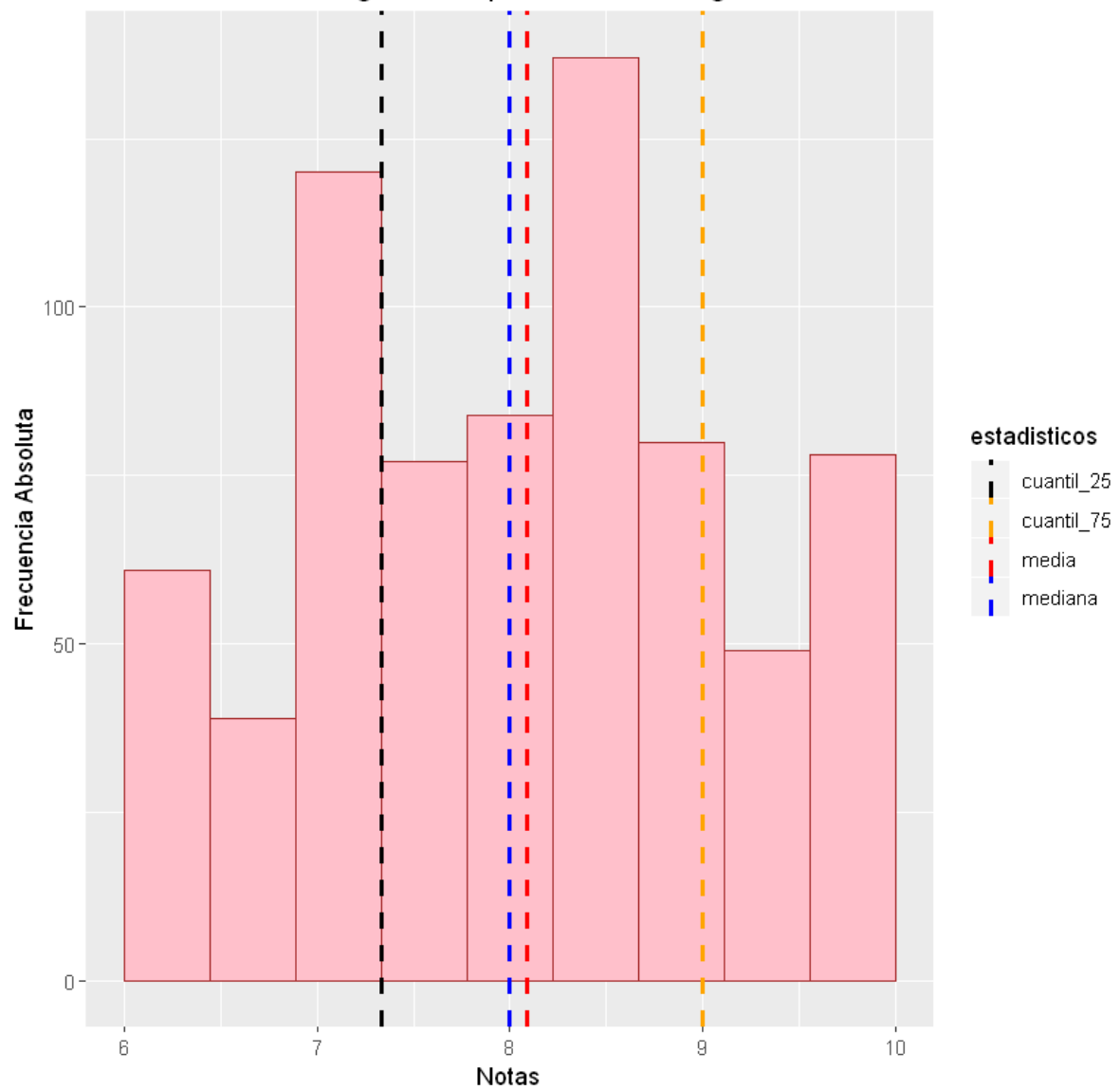


Gráfico 19

Secundaria - Histograma de promedio de Matemática - 1 año.

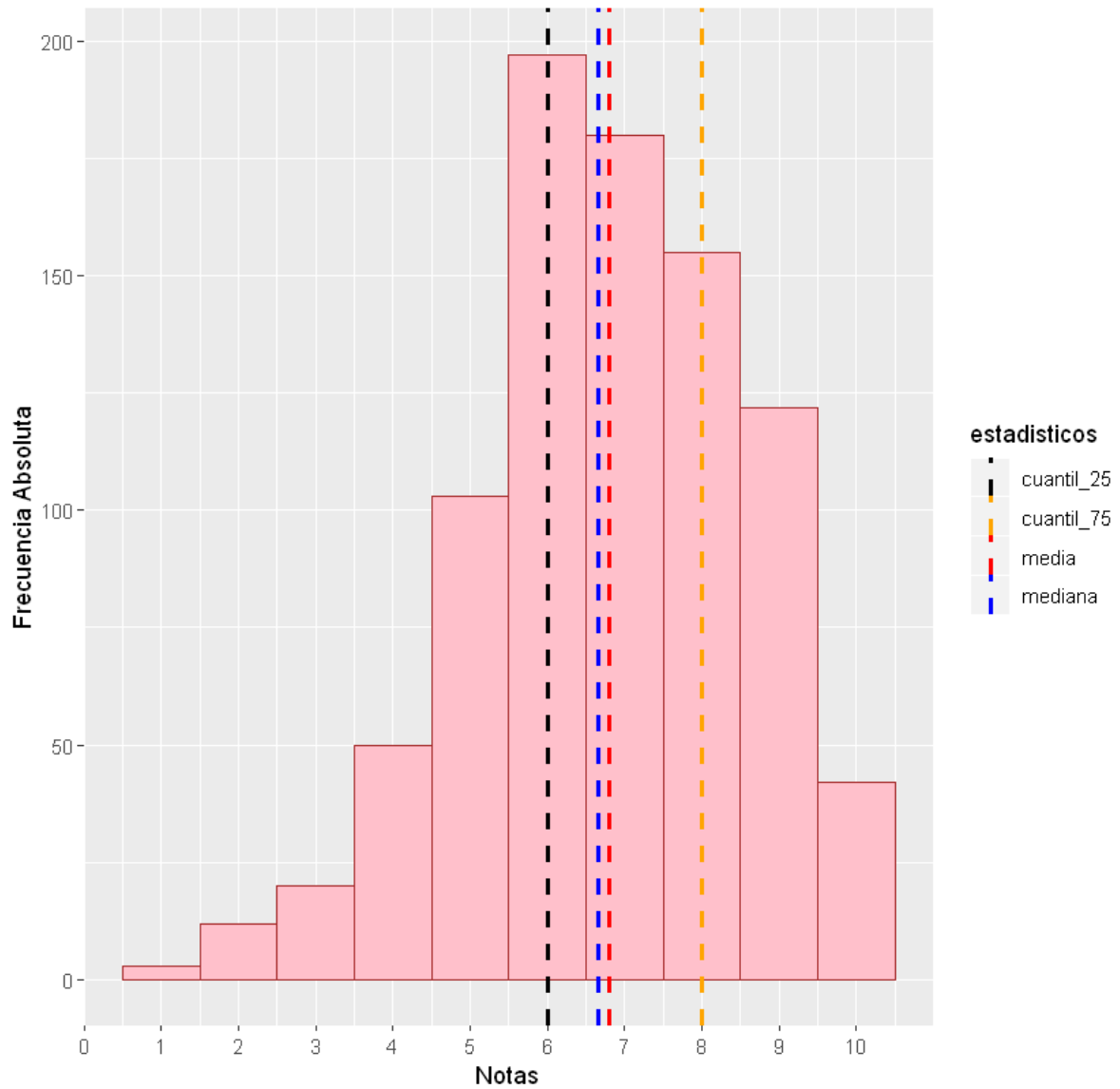


Gráfico 20

Secundaria - Histograma de promedio de Lengua y Literatura - 1 año.

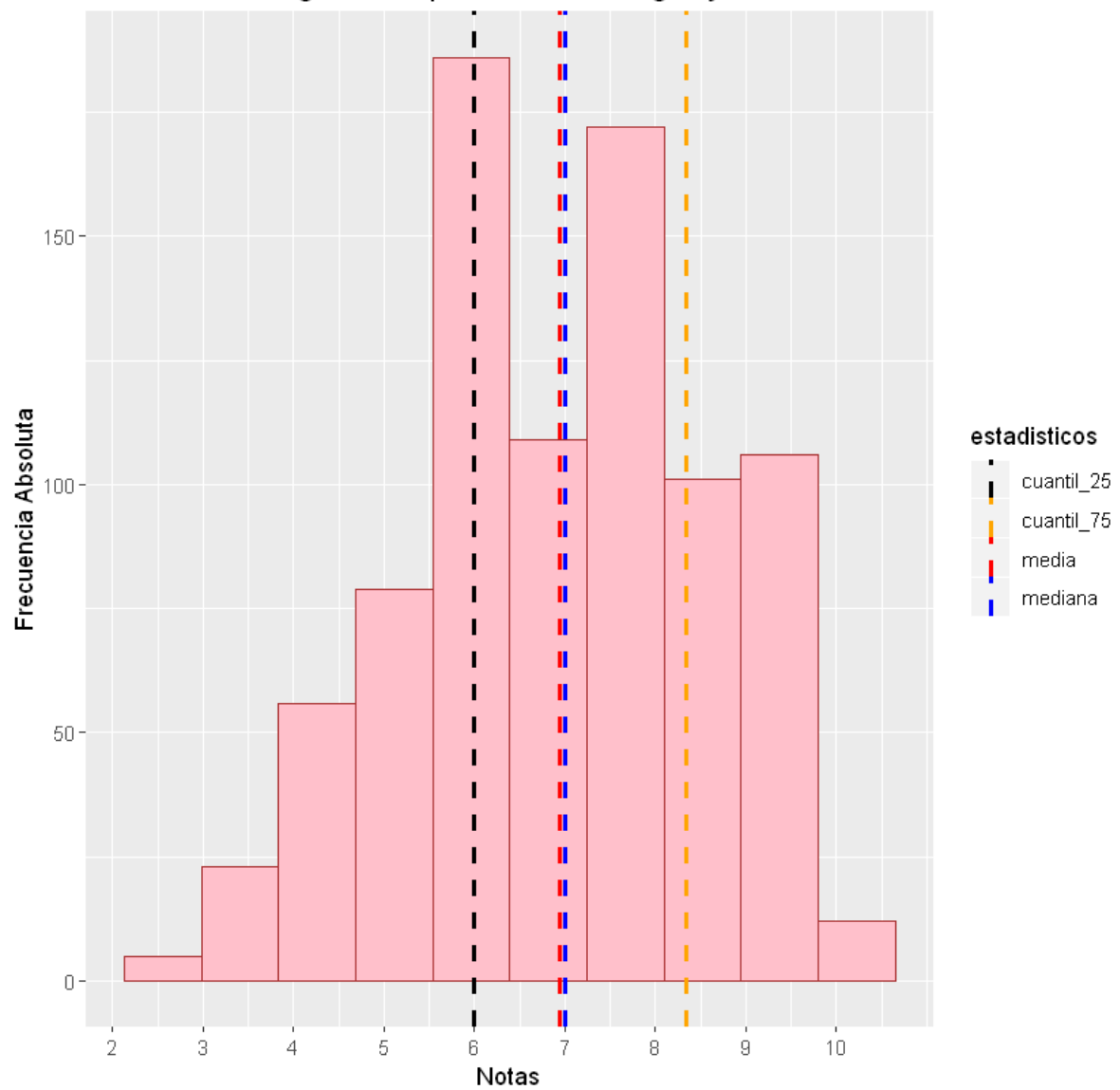


Gráfico 21

Secundaria - Histograma de promedio de Matemática - 2 año.

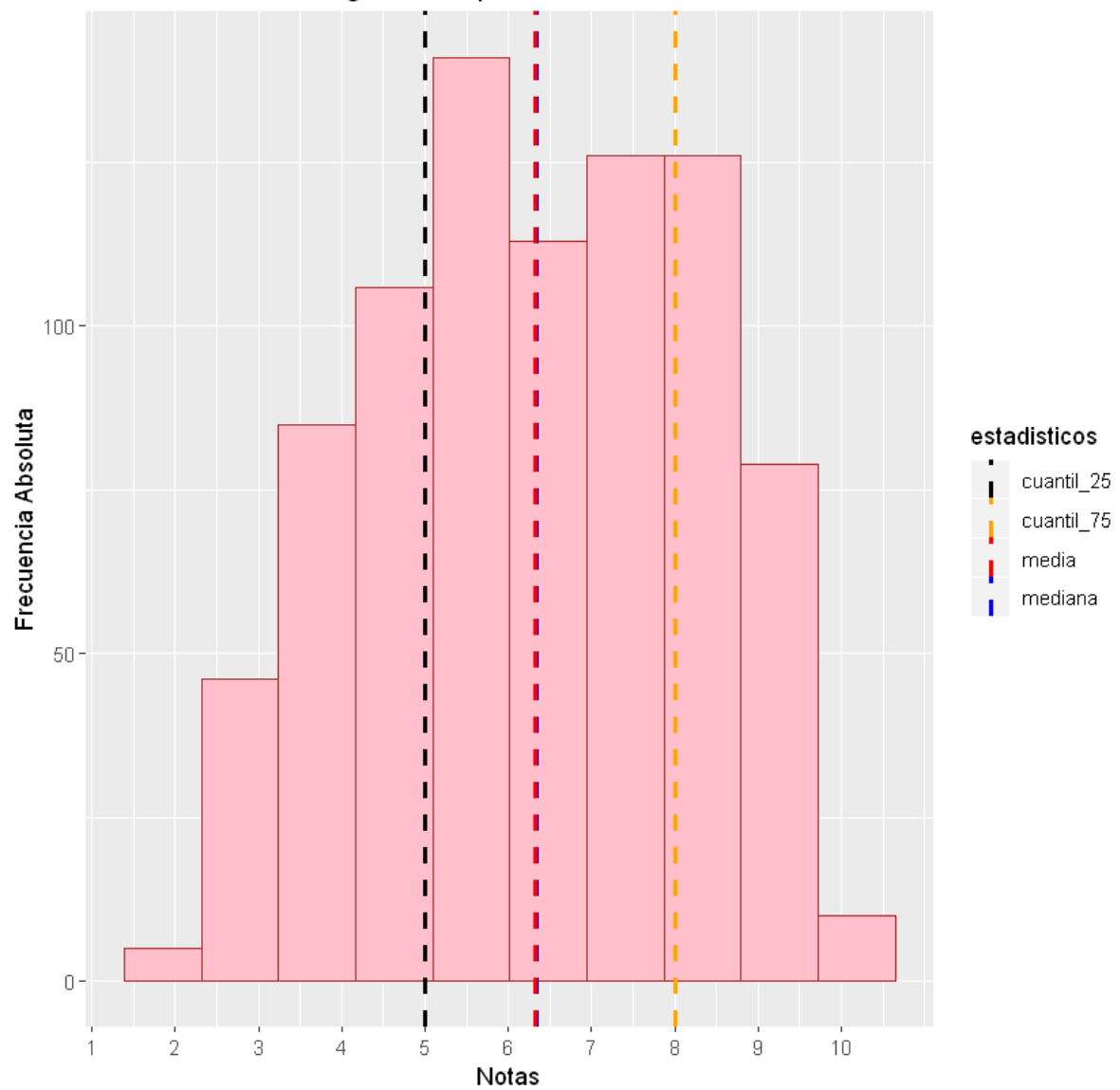


Gráfico 22

Secundaria - Histograma de promedio de Lengua y Literatura - 2 año.

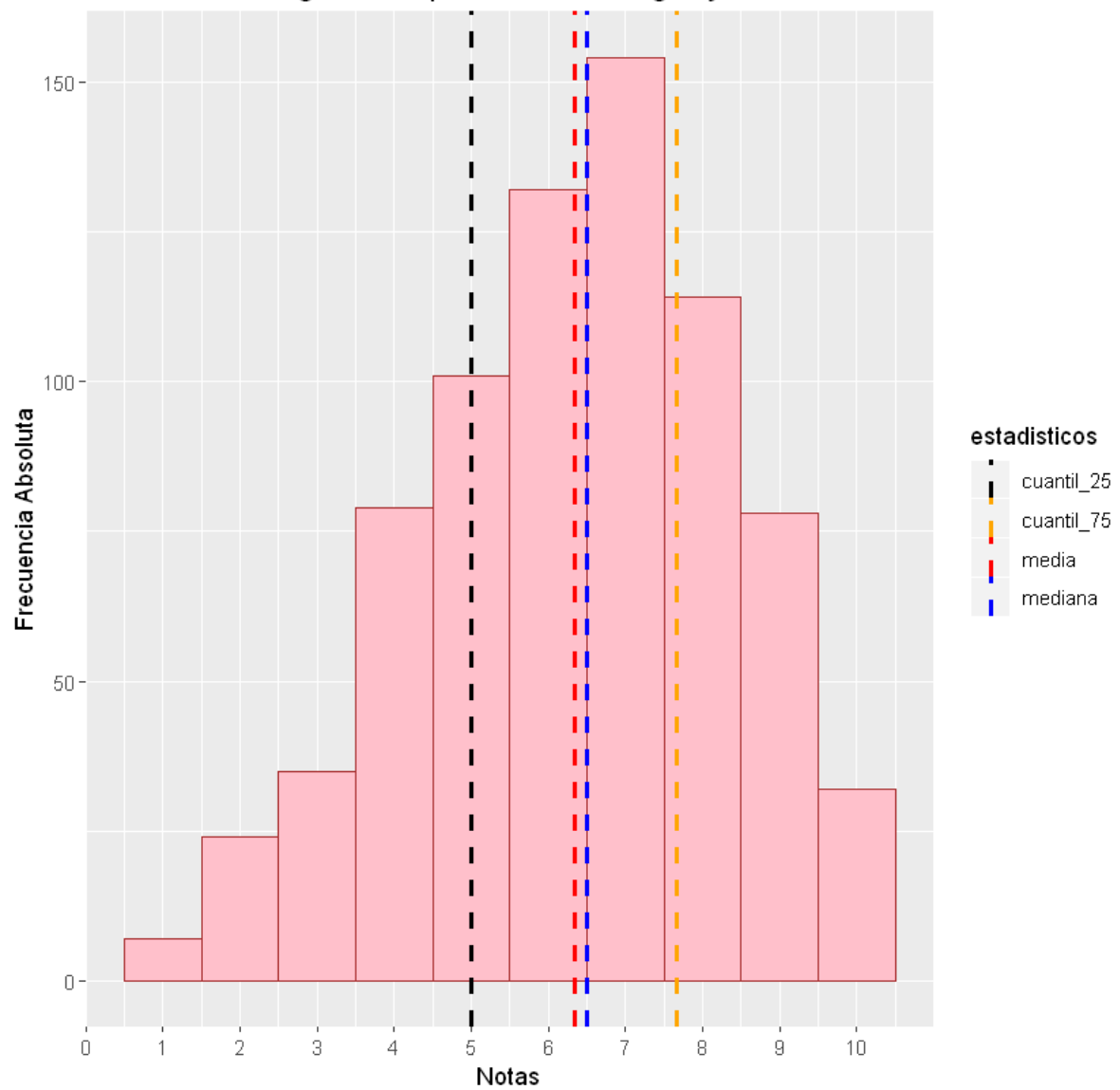


Gráfico 23

Secundaria - Histograma de promedio de Matemática - 3 año.

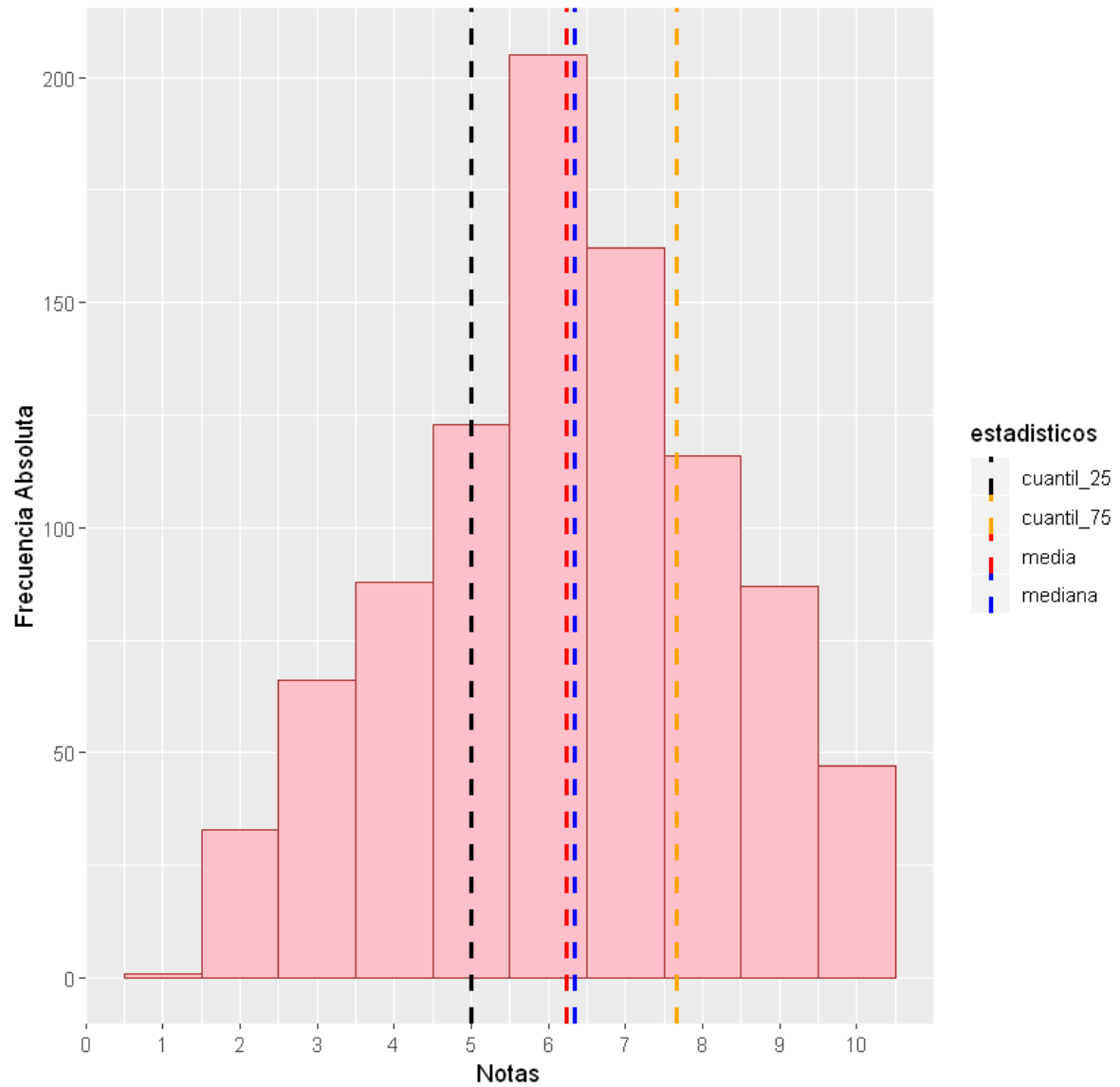


Gráfico 24

Secundaria - Histograma de promedio de Lengua y Literatura - 3 año.

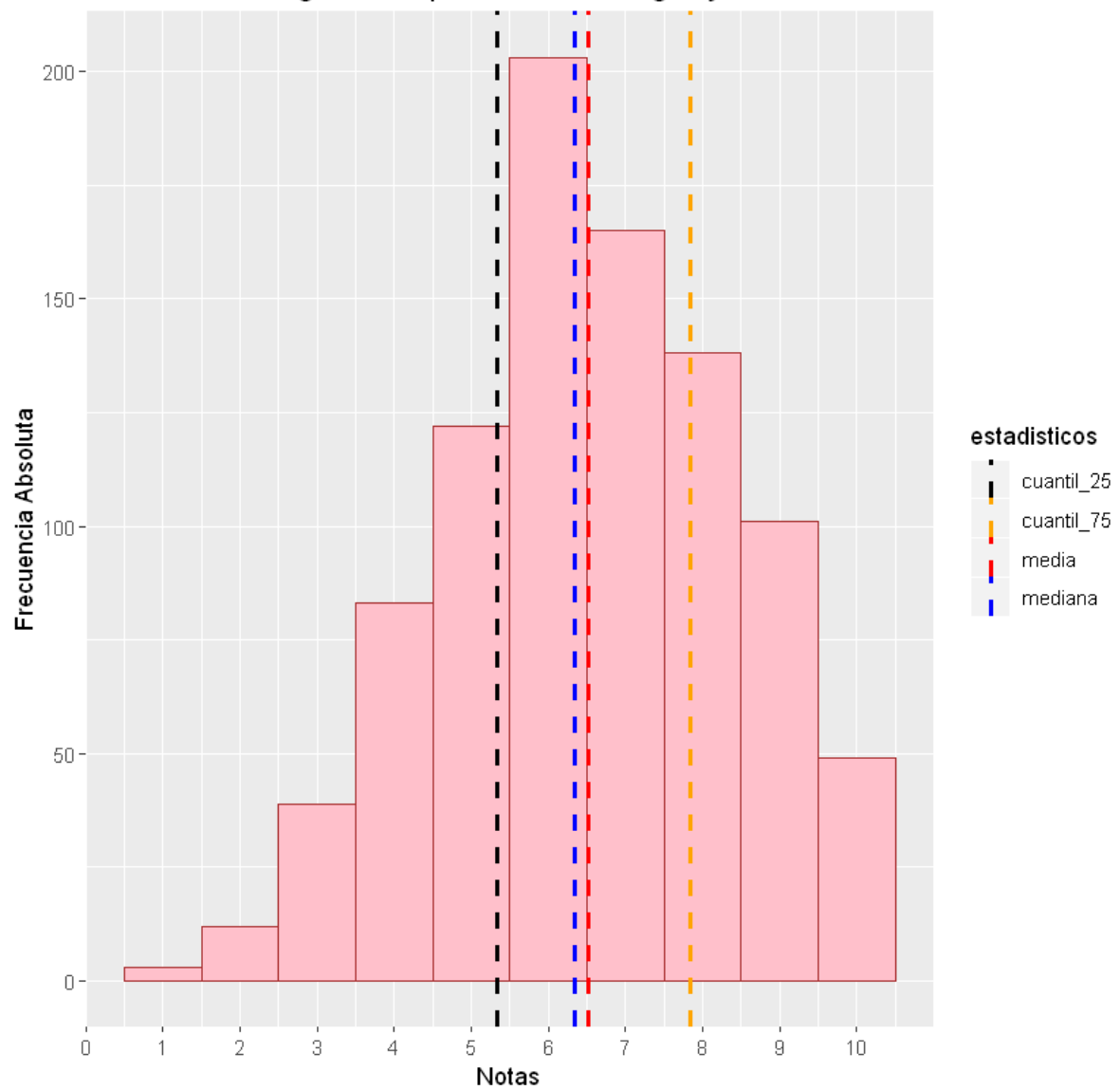


Gráfico 25

Secundaria - Histograma de promedio de Matemática - 4 año.

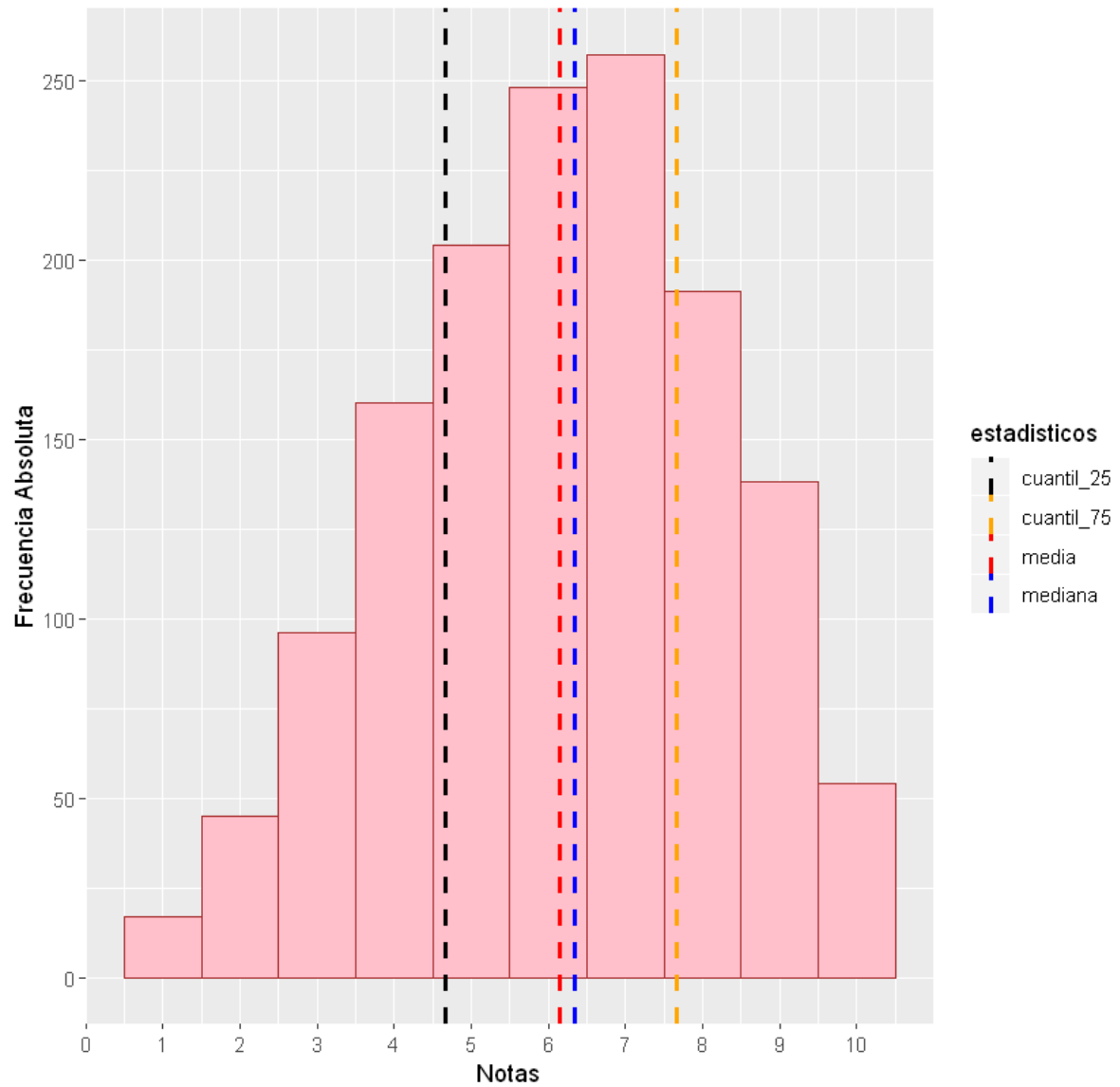


Gráfico 26

Secundaria - Histograma de promedio de Lengua y Literatura - 4 año.

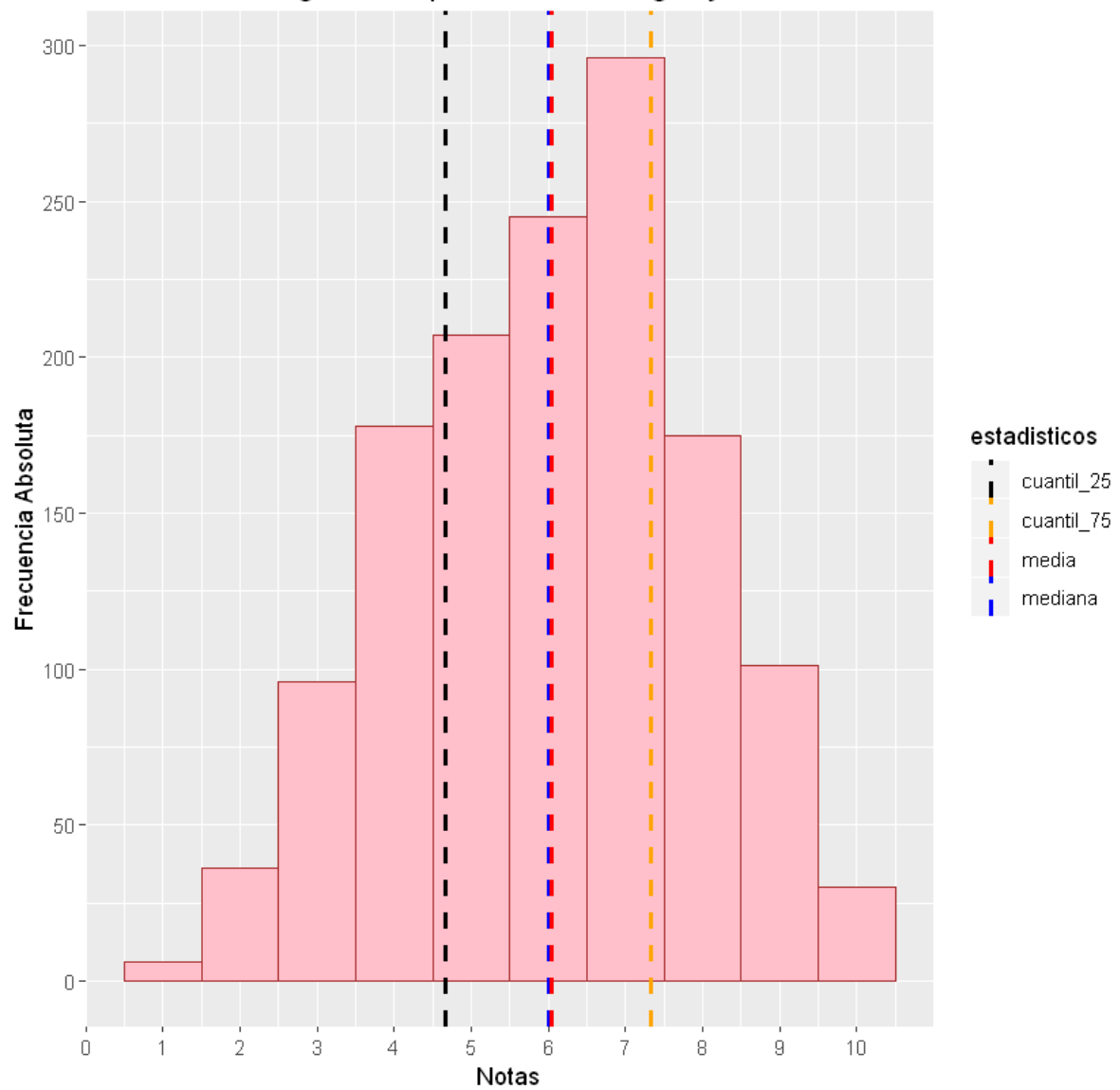


Gráfico 27

Secundaria - Histograma de promedio de Matemática - 5 año.

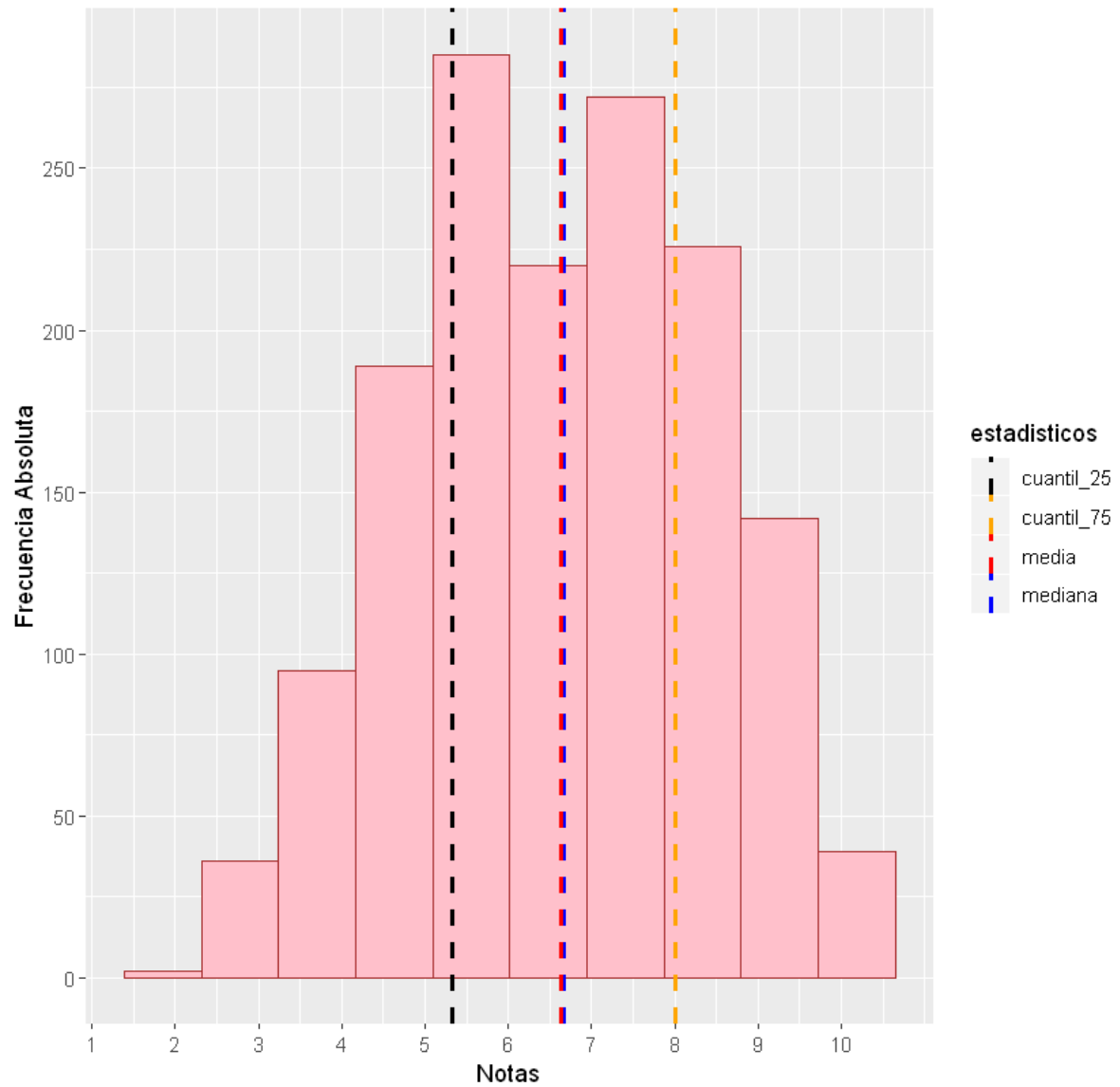


Gráfico 28

Secundaria - Histograma de promedio de Lengua y Literatura - 5 año.

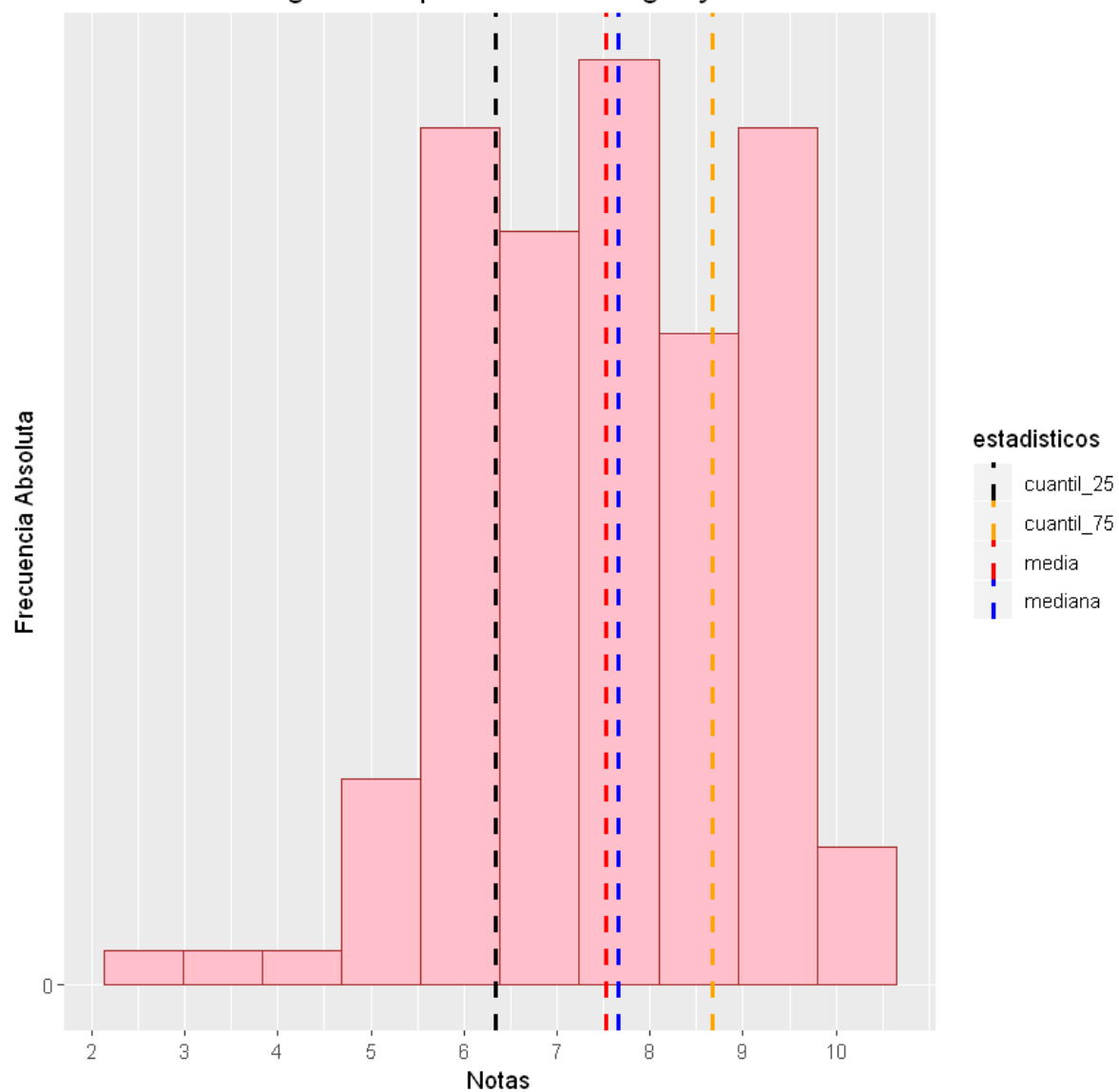


Gráfico 29

Secundaria - Histograma de promedio de Matemática - 6 año.

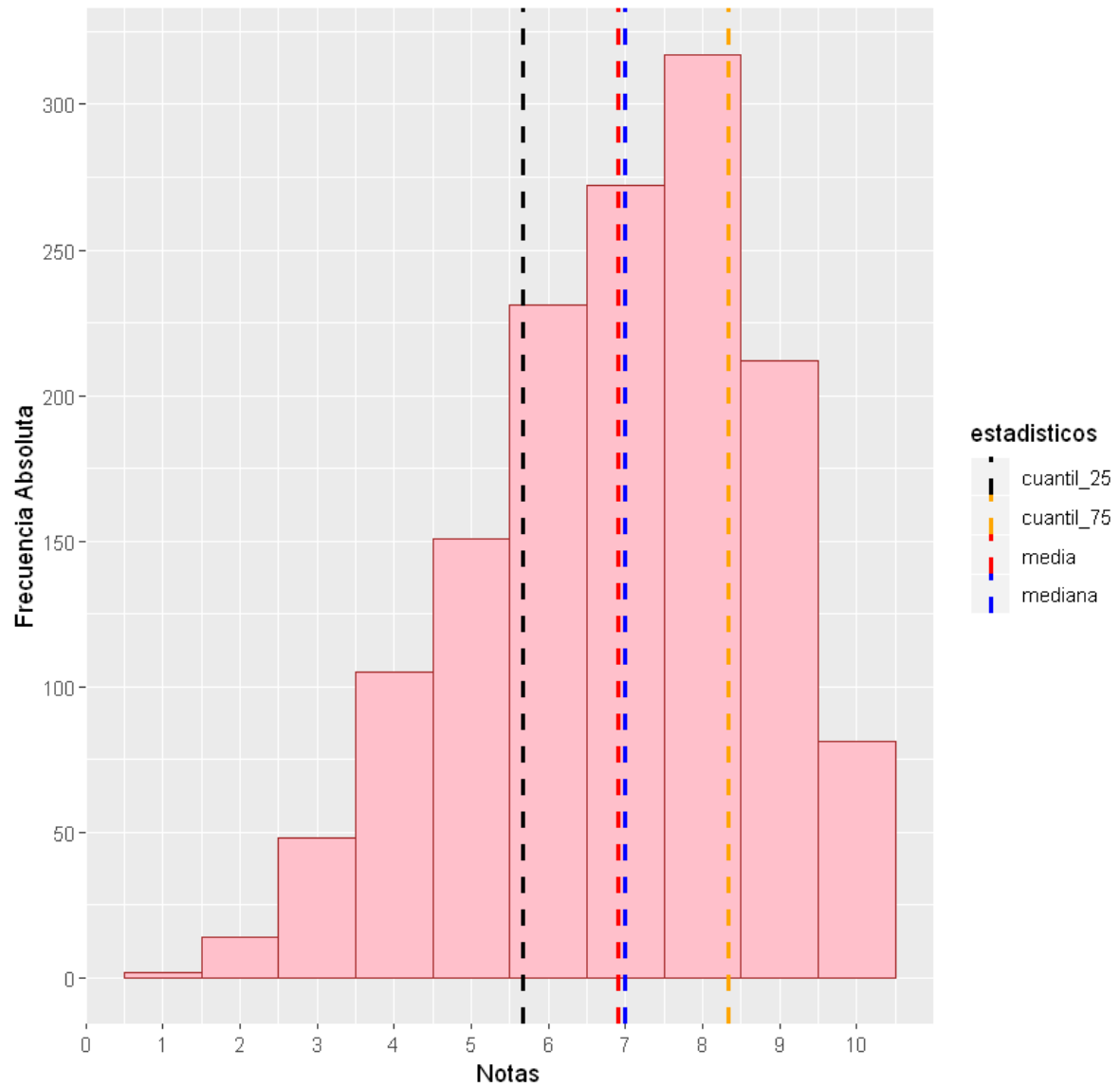


Gráfico 30

Secundaria - Histograma de promedio de Lengua y Literatura - 6 año.

