M1.2 Datos Faltantes y Outliers

Integrantes

A01068244 - Jared Andrés Silva Villa
A00227869 - Paola Félix Torres

Fecha: 15 de Agosto del 2024

Índice

1.Identificar el porcentaje de datos faltantes.

El porcentaje de datos faltantes en la variable de **absences** es de **7.05%**, mientras que en la variable de **traveltime** es de **5.61%**.

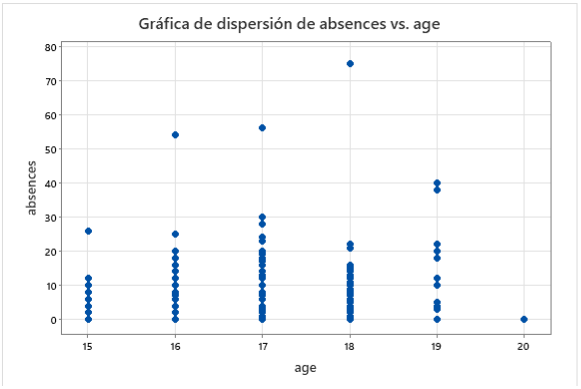
2.Identificar el mecanismo que ocasiona datos faltantes (MCAR, MAR, NMAR)}

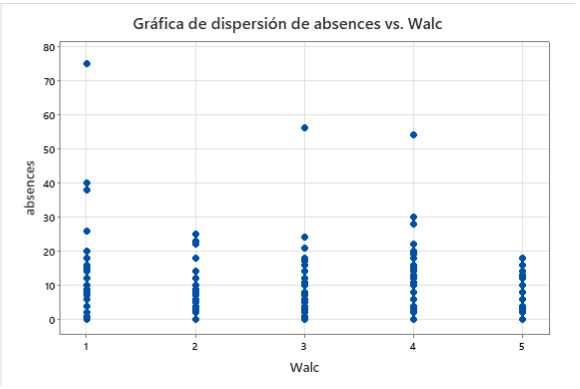
Concluimos que, como no hay muchos datos faltantes, lo más probable es que el mecanismo sea **MCAR.** También hicimos un análisis de correlación y vimos que las variables que más se relacionan con absences son age y walc, y para traveltime son salc y walc. Sin embargo, al ver las gráficas de dispersión, no encontramos nada raro que nos haga pensar en otro tipo de mecanismo distinto a MCAR.

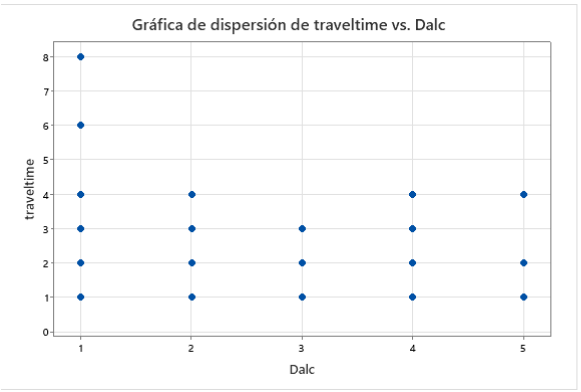
Correlaciones

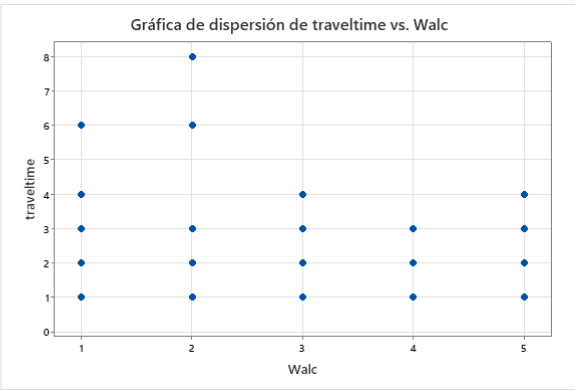
	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout
Medu	-0.164								
Fedu	-0.169	0.631							
traveltime	0.112	-0.141	-0.114						
studytime	0.044	0.051	0.053	-0.040					
failures	0.244	-0.237	-0.255	0.093	-0.114				
famrel	0.054	-0.004	-0.037	0.032	0.006	-0.044			
freetime	0.016	0.031	-0.027	-0.014	-0.181	0.092	0.151		
goout	0.127	0.064	0.024	0.008	-0.050	0.125	0.065	0.285	
Dalc	0.338	-0.037	-0.044	0.118	-0.063	0.172	-0.059	0.176	0.206
Walc	0.117	-0.047	-0.017	0.121	-0.154	0.142	-0.113	0.148	0.420
health	-0.062	-0.047	0.034	-0.004	-0.049	0.066	0.094	0.076	-0.010
absences	0.173	0.103	0.030	-0.040	-0.064	0.013	-0.044	-0.062	0.023

	Dalc	Walc	health
Medu			
Fedu			
traveltime			
studytime			
failures			
famrel			
freetime			
goout			
Dalc			
Walc	0.598		
health	0.057	0.092	
absences	0.077	0.117	-0.020

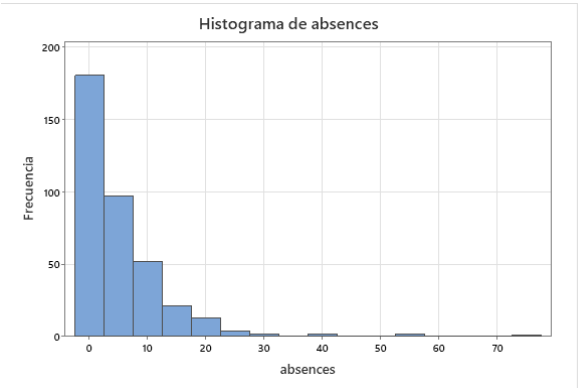


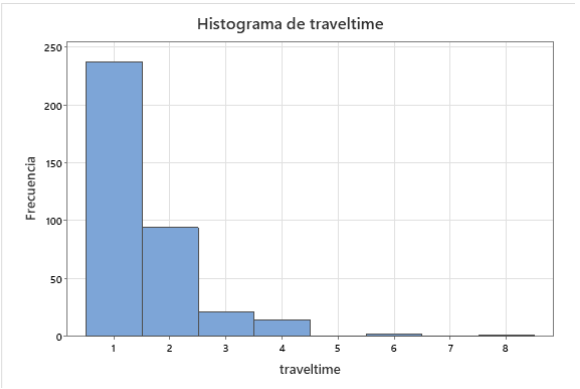






3.Obtener estadísticas descriptivas de los datos (histograma, media, desviación estándar, mediana, moda, etc).





Estadísticas

Variable	Conteo total	N	N*	Media	Error estándar de la media	Desv.Est.	Mínimo	Q1	Mediana
absences	395	374	21	5.543	0.418	8.089	0.000	0.000	3.500
traveltime	395	369	26	1.5285	0.0470	0.9028	1.0000	1.0000	1.0000

Variable	Q3	Máximo
absences	8.000	75.000
traveltime	2.0000	8.0000

Estadísticas

Variable	Conteo total	N	N*	Media	Error estándar de la media	Desv.Est.	Mínimo	Q1	Mediana
age	395	395	0	16.696	0.0642	1.276	15.000	16.000	17.000
Medu	395	395	0	2.7494	0.0551	1.0947	0.0000	2.0000	3.0000
Fedu	395	363	32	2.5207	0.0578	1.1007	0.0000	2.0000	2.0000
traveltime	395	369	26	1.5285	0.0470	0.9028	1.0000	1.0000	1.0000
studytime	395	395	0	2.1595	0.0634	1.2594	1.0000	1.0000	2.0000
failures	395	395	0	0.3342	0.0374	0.7437	0.0000	0.0000	0.0000
famrel	395	395	0	3.9443	0.0451	0.8967	1.0000	4.0000	4.0000
freetime	395	395	0	3.2354	0.0503	0.9989	1.0000	3.0000	3.0000
goout	395	395	0	3.1089	0.0560	1.1133	1.0000	2.0000	3.0000
Dalc	395	324	71	1.3580	0.0446	0.8034	1.0000	1.0000	1.0000
Walc	395	395	0	2.2911	0.0648	1.2879	1.0000	1.0000	2.0000
health	395	395	0	3.5544	0.0700	1.3903	1.0000	3.0000	4.0000
absences	395	374	21	5.543	0.418	8.089	0.000	0.000	3.500

Variable	Q3	Máximo	Moda	N para moda
age	18.000	22.000	16	104
Medu	4.0000	4.0000	4	131
Fedu	3.0000	4.0000	2	102
traveltime	2.0000	8.0000	1	237
studytime	2.0000	12.0000	2	193
failures	0.0000	3.0000	0	312
famrel	5.0000	5.0000	4	195
freetime	4.0000	5.0000	3	157
goout	4.0000	5.0000	3	130
Dalc	1.0000	5.0000	1	253
Walc	3.0000	5.0000	1	151
health	5.0000	5.0000	5	146
absences	8.000	75.000	0	115

4.Utilizar el método de imputación adecuado para cada una de las variables con datos faltantes.

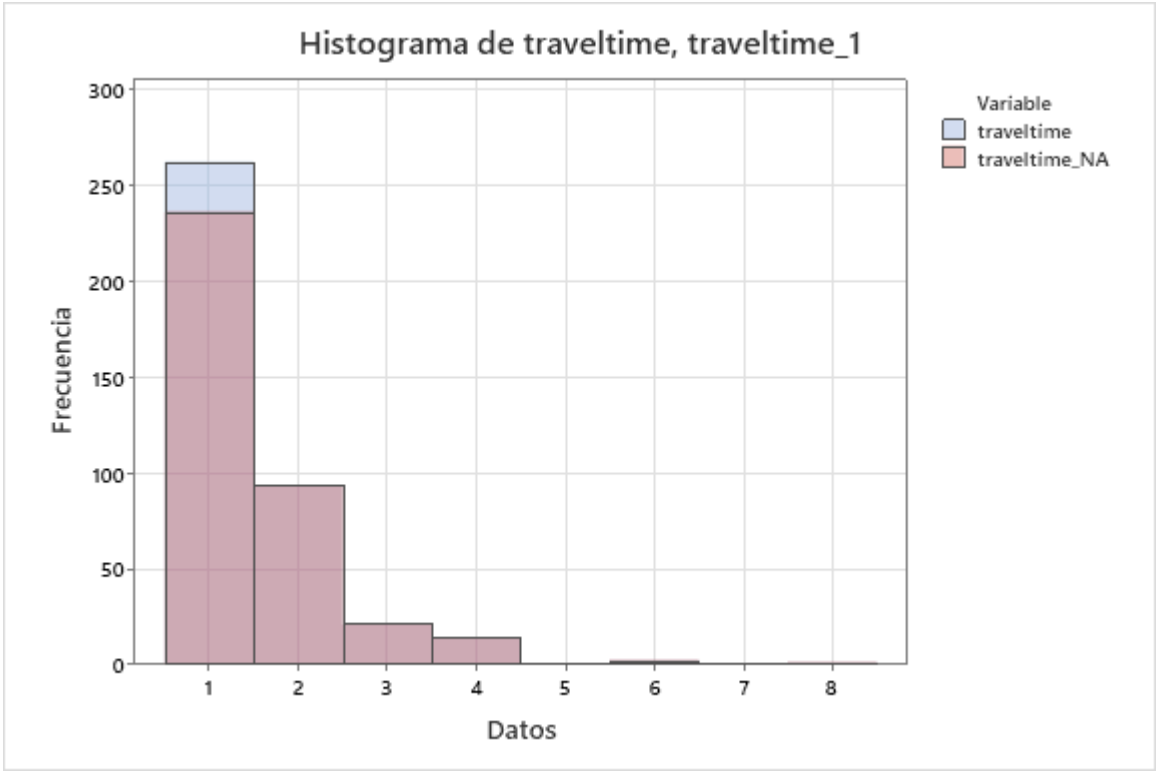
Tomando el cuenta el valor de asimetría, podemos darnos cuenta que las dos variables son asimétricas, segadas hacías la derecha, por lo que la imputación simple adecuada sería utilizar la **Mediana** en ambos casos.

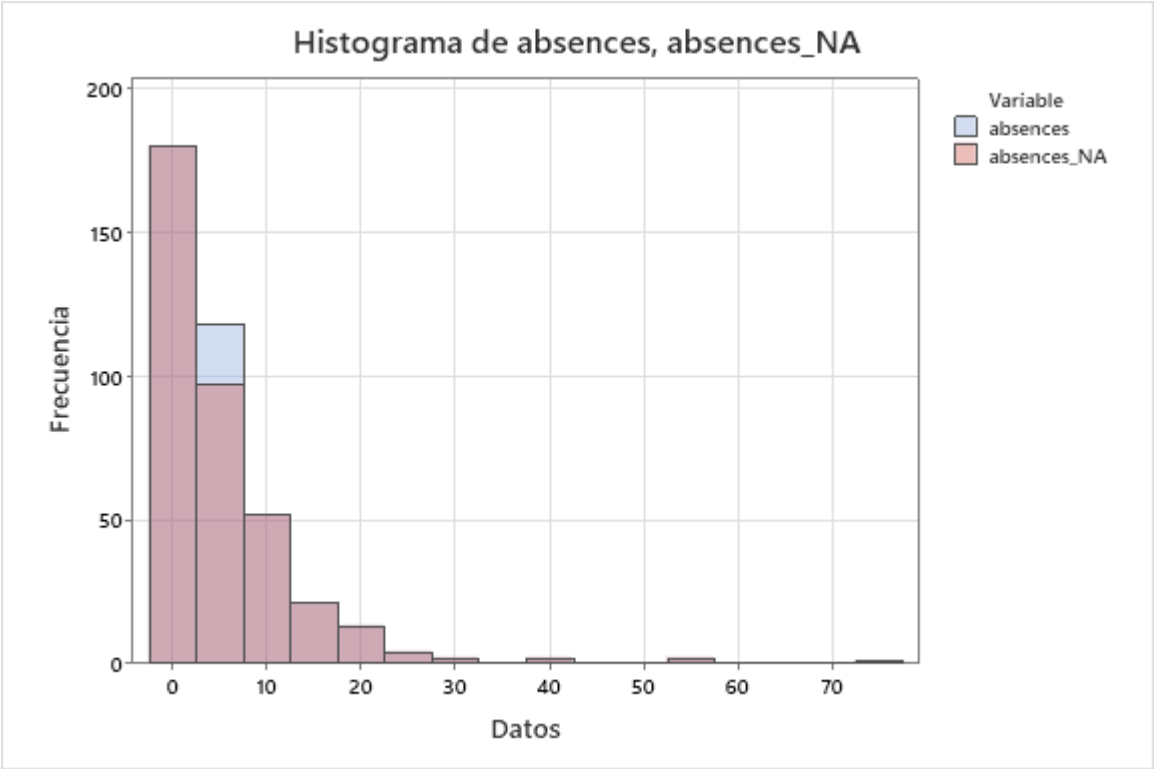
Valor de asimetría
Absences = 3.78
Traveltime = 2.61

Estadísticas

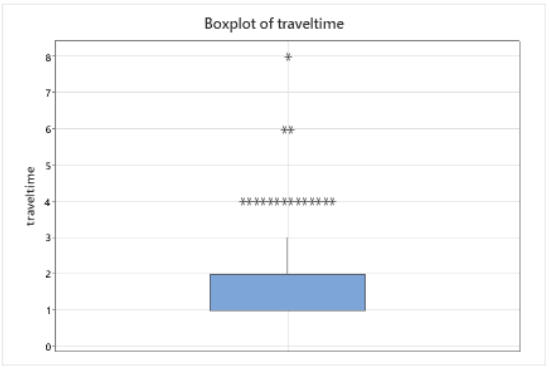
Variable	Conteo total	N	N*	Media	Error estándar de la media	Desv.Est.	Mínimo	Q1	Mediana
absences	395	374	21	5.543	0.418	8.089	0.000	0.000	3.500
traveltime	395	369	26	1.5285	0.0470	0.9028	1.0000	1.0000	1.0000

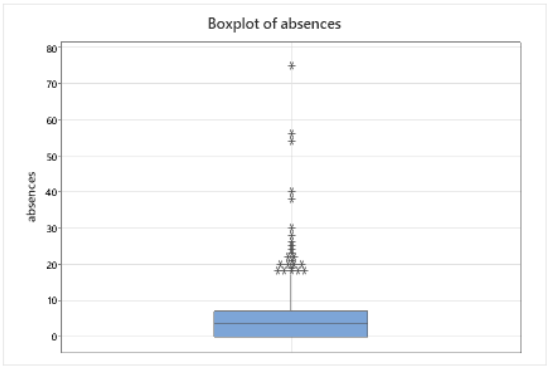
Variable	Q3	Máximo	Modo	N para moda	Asimetría
absences	8.000	75.000	0	115	3.78
traveltime	2.0000	8.0000	1	237	2.61





5.Realizar un boxplot e interpretarlo.





Ambos boxplots muestran que la mayoría de los estudiantes tienen un tiempo de transporte corto y pocas ausencias. En el boxplot de traveltime, la mediana está en el nivel 1, con la mayoría de los estudiantes en un rango de 1 a 2, pero hay outliers que muestran tiempos de transporte significativamente más largos, alcanzando el nivel 8 (lo cual técnicamente no debería ser posible). En el boxplot de absences, la mediana es de 4 días de ausencia, con la mayoría de los estudiantes faltando entre 0 y 10 días. Sin embargo, también hay outliers que muestran estudiantes con un número mucho mayor de ausencias, algunos superando los 70 días.