1. Identificar el porcentaje de datos faltantes.

   El porcentaje de datos faltantes en traveltime es: 6.5823% y el porcentaje en absences es: 5.3164%.

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

Correlations

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

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

En Traveltime debido a la baja correlación con la mayoría de las variables, podría significar que los valores faltantes no tienen nada que ver o no dependen de las otras variables, por lo tanto, se podría decir que esto ocurre al azar. (MCAR).

En Absences puede ser un poco más confuso. Aunque sí hay una ligera correlación con las variables que están relacionadas con el consumo de alcohol, la correlación es baja. Por lo que los datos faltantes también se deben al azar. (MCAR).

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

Descriptive Statistics: ...

WORKSHEET 2

Descriptive Statistics: traveltime, absences

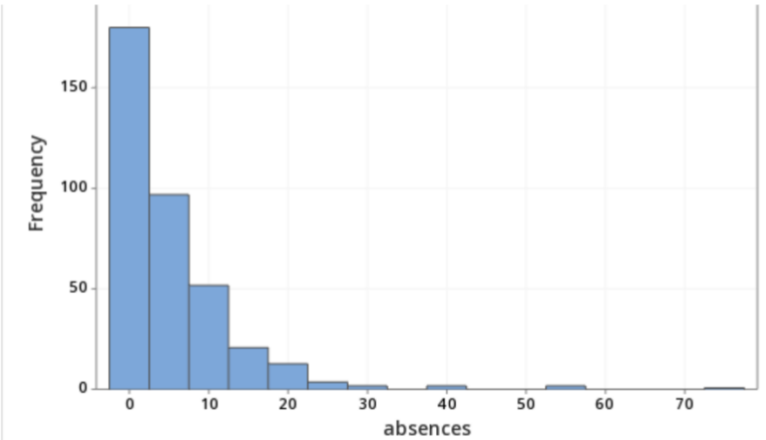
Statistics

Variable	Total Count	N	N*	Percent	Mean	SE Mean	StDev	Variance	CoefVar
traveltime	395	369	26	93.4177	1.5285	0.0470	0.9028	0.8151	59.07
absences	395	374	21	94.6835	5.543	0.418	8.089	65.434	145.94

Variable	Minimum	Median	Maximum	Range	Mode	N for Mode
traveltime	1.0000	1.0000	8.0000	7.0000	1	237
absences	0.000	3.500	75.000	75.000	0	115

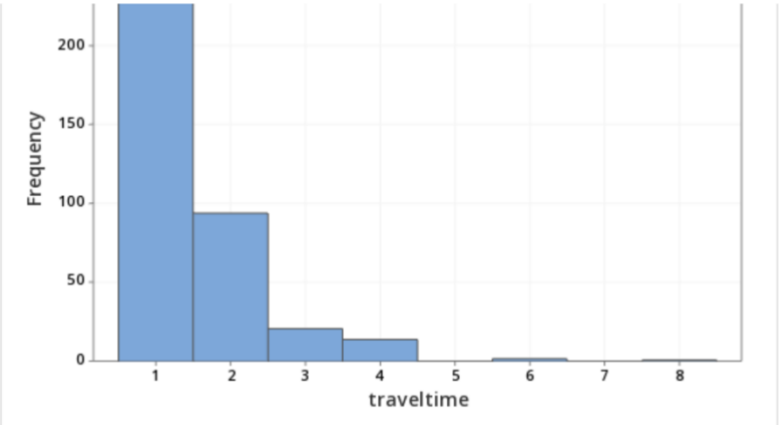
WORKSHEET 2

Histogram of absences



WORKSHEET 2

Histogram of traveltime



WORKSHEET 1

Descriptive Statistics: age, Medu, Fedu, traveltime, studytime, failures, famrel, ...

Statistics

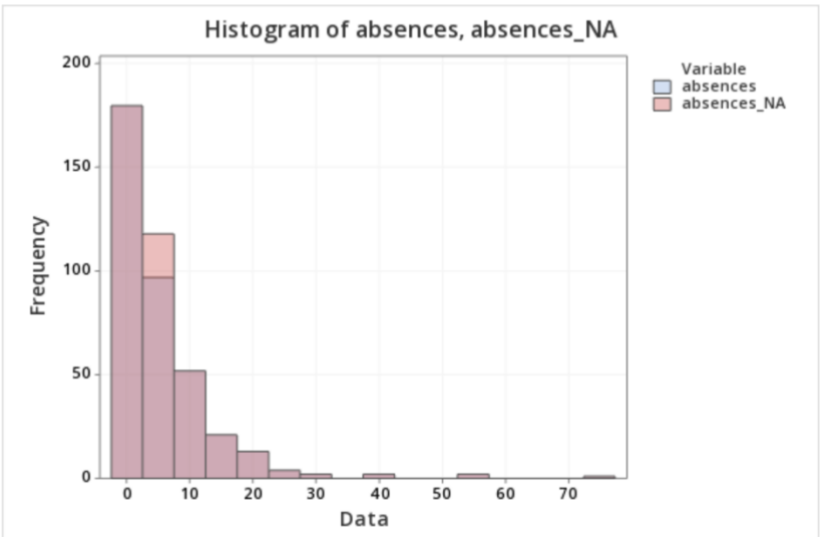
Variable	Total Count	N	N*	Percent	Mean	SE Mean	StDev	Variance	CoefVar
age	395	395	0	100.000	16.696	0.0642	1.276	1.628	7.64
Medu	395	395	0	100.000	2.7494	0.0551	1.0947	1.1984	39.82
Fedu	395	363	32	91.899	2.5207	0.0578	1.1007	1.2116	43.67
traveltime	395	369	26	93.418	1.5285	0.0470	0.9028	0.8151	59.07
studytime	395	395	0	100.000	2.1595	0.0634	1.2594	1.5862	58.32
failures	395	395	0	100.000	0.3342	0.0374	0.7437	0.5530	222.53
famrel	395	395	0	100.000	3.9443	0.0451	0.8967	0.8040	22.73
freetime	395	395	0	100.000	3.2354	0.0503	0.9989	0.9977	30.87
goout	395	395	0	100.000	3.1089	0.0560	1.1133	1.2394	35.81
Dalc	395	324	71	82.025	1.3580	0.0446	0.8034	0.6454	59.16
Walc	395	395	0	100.000	2.2911	0.0648	1.2879	1.6587	56.21
health	395	395	0	100.000	3.5544	0.0700	1.3903	1.9329	39.11
absences	395	374	21	94.684	5.543	0.418	8.089	65.434	145.94

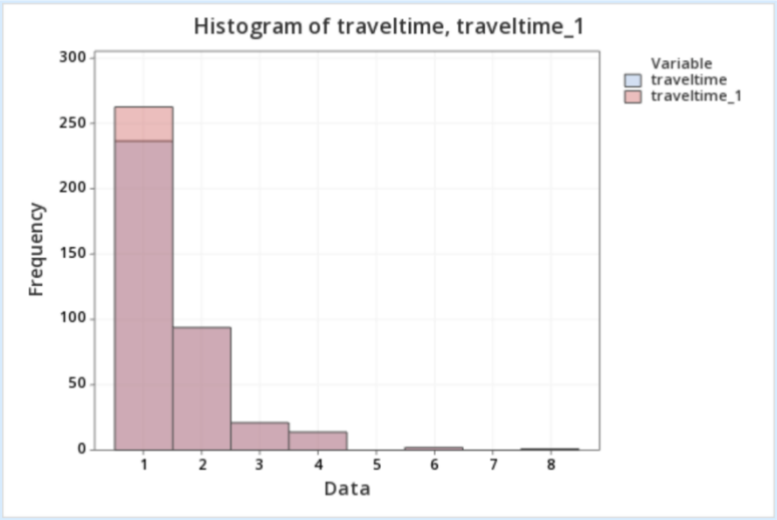
Variable	Minimum	Median	Maximum	Range	Mode	N for Mode	Skewness
age	15.000	17.000	22.000	7.000	16	104	0.47
Medu	0.0000	3.0000	4.0000	4.0000	4	131	-0.32
Fedu	0.0000	2.0000	4.0000	4.0000	2	102	-0.04
traveltime	1.0000	1.0000	8.0000	7.0000	1	237	2.61
studytime	1.0000	2.0000	12.0000	11.0000	2	193	3.36
failures	0.0000	0.0000	3.0000	3.0000	0	312	2.39
famrel	1.0000	4.0000	5.0000	4.0000	4	195	-0.95
freetime	1.0000	3.0000	5.0000	4.0000	3	157	-0.16
goout	1.0000	3.0000	5.0000	4.0000	3	130	0.12
Dalc	1.0000	1.0000	5.0000	4.0000	1	253	2.65
Walc	1.0000	2.0000	5.0000	4.0000	1	151	0.61
health	1.0000	4.0000	5.0000	4.0000	5	146	-0.49
absences	0.000	3.500	75.000	75.000	0	115	3.78

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

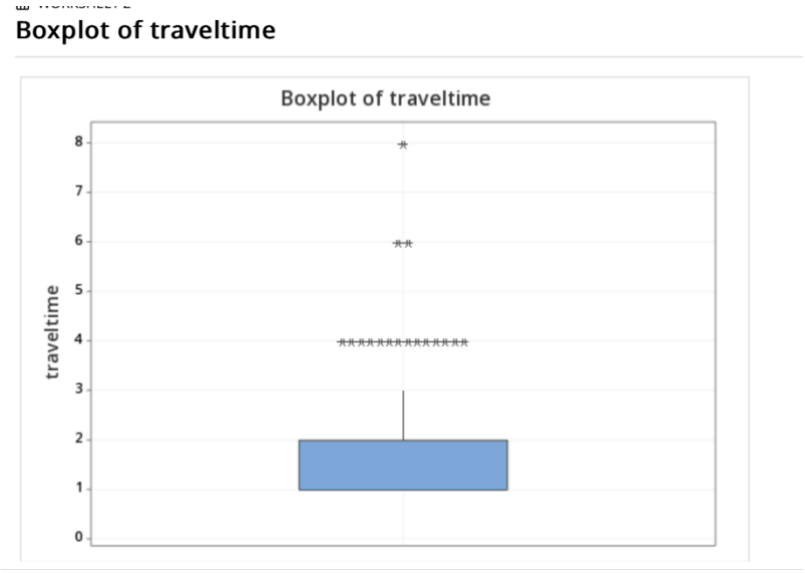
Por los valores de asimetría resultantes en la descripción de estadísticas, que son 2.61 para la variable traveltime y 3.78 para absences, y tomando en cuenta que son mayores a 0, veo el histograma y percibo que tiene sesgo hacia la derecha, por lo tanto, el método adecuado de imputación sería la mediana.

Histogram of absences, absences_NA





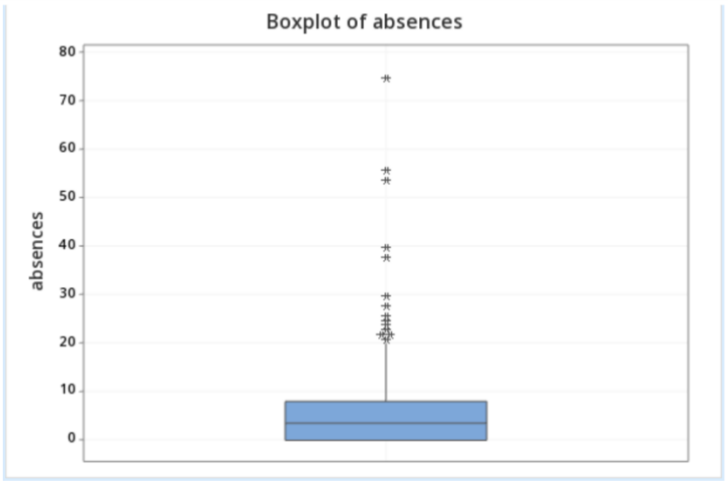
5. Realizar un boxplot e interpretarlo.



En cuanto a la distribución, se puede percibir que la mayoría de los estudiantes tienen tiempos de viaje cortos con la mayoría estando en 1. Algunos estudiantes tienen tiempos de viaje más largos, que son casos menos frecuentes.

La distribución es asimétrica hacia la derecha con cola extendida hacia tiempos de viaje más largos, los cuales se muestran en valores atípicos (*).

Boxplot of absences



En este boxplot, se puede ver que la mayoría de los estudiantes tienen un número bajo de faltas, con la mayoría entre 0 a 8.
Hay varios valores atípicos que nos indican que algunos de los estudiantes tienen valores de faltas más altos.
La distribución asimétrica va hacia la derecha, con cola extendida hacia los valores de faltas más altos, los cuales se reflejan en los valores atípicos (*).