M1.2 Datos Faltantes y Outliers

Participantes:

Carlos Dhali Tejeda Tapia – A00344820

Enrique Mora Navarro - A01635459

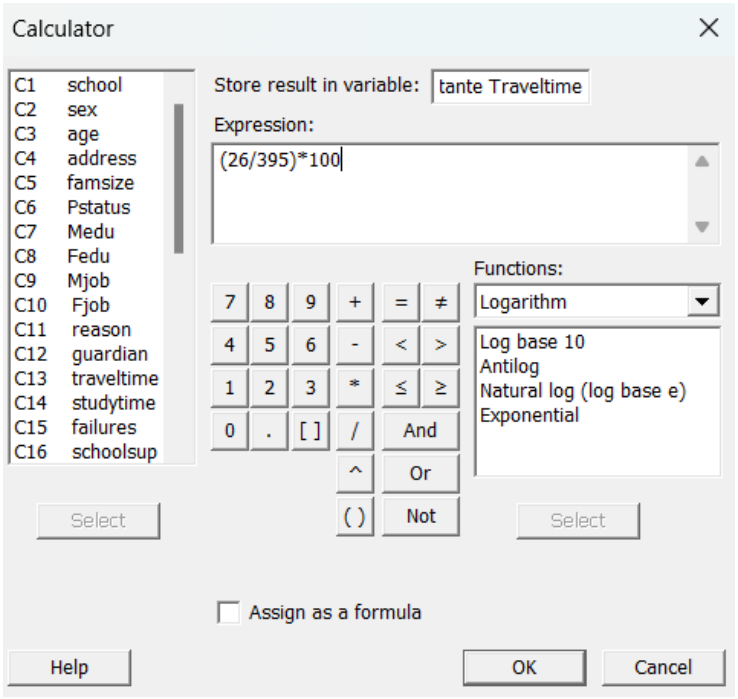
1.Identificar el porcentaje de datos faltantes.

C38	C39
%Faltante Traveltime	%Faltante absences
6.58228	5.31646

Para calcular el porcentaje de los datos faltantes de cada columna utilicé esta fórmula:

$$\text{Porcentaje de valores faltantes} = \left(\frac{\text{Número de valores faltantes}}{\text{Número total de filas en la columna}} \right) \times 100$$

Calculo en Minitab:



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

Para identificar el mecanismo decidimos basarnos en una prueba de correlación entre las variables numéricas:

Correlations

	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

Viendo que en la prueba de correlación las variables de traveltime y absences no muestra que exista una correlación notable entre estas variables con las demás. A pesar de que las pruebas de correlación no son un dato absoluto si se puede usar para identificar el mecanismo que está ocasionando los datos faltantes y en este caso y basándonos en las pruebas de correlación, nosotros creemos que los datos faltantes de estas 2 variables fueron observados totalmente al azar. **MCAR**

- **MCAR** (Missing completely at Random): la probabilidad de observar el dato faltante es completamente al azar.
  - *Ej: Un encuestado se salta accidentalmente una pregunta por descuido.*

¿Cómo identificar si es MCAR, MAR o NMAR?

3. **Pruebas de Correlación**: ***Correlaciones*** entre una variable dummy (si hay dato faltante o no) y otras variables. Si hay correlación puede ser MAR; si no, puede ser MCAR. No distinge entre MAR y NMAR. Además, ausencia de correlación no garantiza MCAR.

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

Statistics

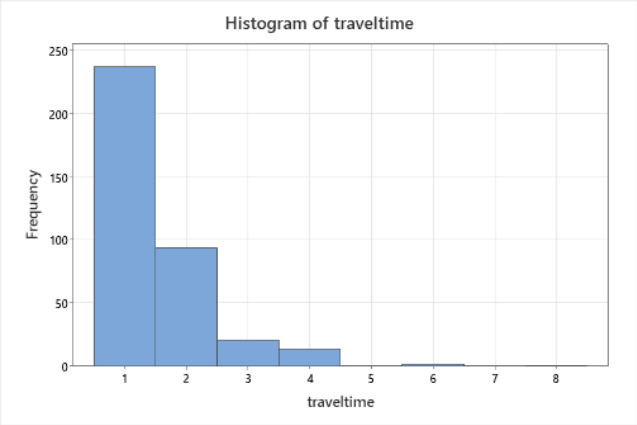
Variable	Total Count	N	N*	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
age	395	395	0	16.696	1.276	15.000	16.000	17.000	18.000	22.000
Medu	395	395	0	2.7494	1.0947	0.0000	2.0000	3.0000	4.0000	4.0000
Fedu	395	363	32	2.5207	1.1007	0.0000	2.0000	2.0000	3.0000	4.0000
traveltime	395	369	26	1.5285	0.9028	1.0000	1.0000	1.0000	2.0000	8.0000
studytime	395	395	0	2.1595	1.2594	1.0000	1.0000	2.0000	2.0000	12.0000
failures	395	395	0	0.3342	0.7437	0.0000	0.0000	0.0000	0.0000	3.0000
famrel	395	395	0	3.9443	0.8967	1.0000	4.0000	4.0000	5.0000	5.0000
freetime	395	395	0	3.2354	0.9989	1.0000	3.0000	3.0000	4.0000	5.0000
goout	395	395	0	3.1089	1.1133	1.0000	2.0000	3.0000	4.0000	5.0000
Dalc	395	324	71	1.3580	0.8034	1.0000	1.0000	1.0000	1.0000	5.0000
Walc	395	395	0	2.2911	1.2879	1.0000	1.0000	2.0000	3.0000	5.0000
health	395	395	0	3.5544	1.3903	1.0000	3.0000	4.0000	5.0000	5.0000
absences	395	374	21	5.543	8.089	0.000	0.000	3.500	8.000	75.000

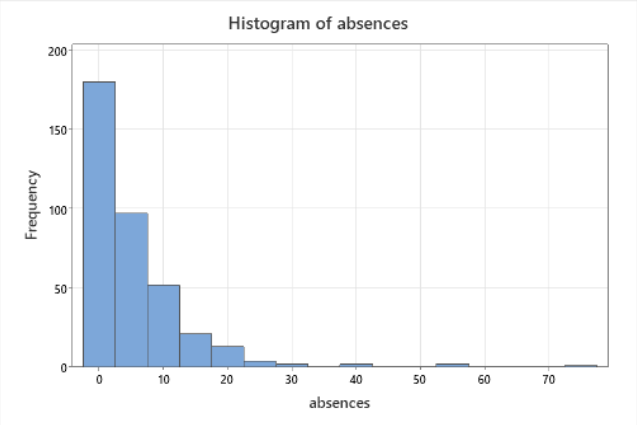
Variable	Mode	N for Mode	Skewness
age	16	104	0.47
Medu	4	131	-0.32
Fedu	2	102	-0.04
traveltime	1	237	2.61
studytime	2	193	3.36
failures	0	312	2.39
famrel	4	195	-0.95
freetime	3	157	-0.16
goout	3	130	0.12
Dalc	1	253	2.65
Walc	1	151	0.61
health	5	146	-0.49
absences	0	115	3.78

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

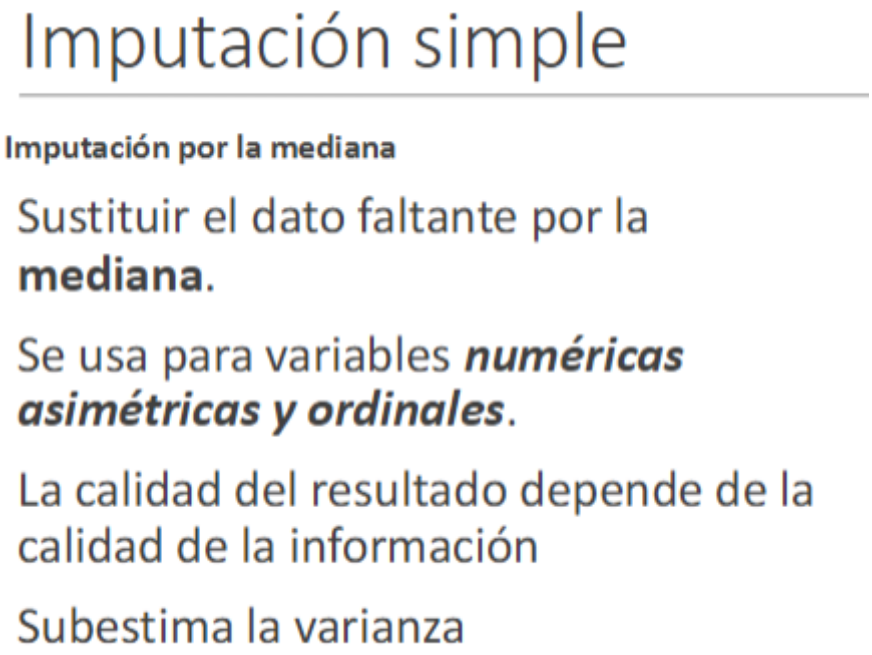
◦**Imputación Simple: Media, Mediana, Moda**

Gracias a los histogramas de las variables que estamos analizando nos dimos cuenta de que nuestras 2 variables son numéricas y asimétricas ya que la mayoría de sus datos se encuentran en los datos iniciales.





Sabiendo que nuestras variables son numéricas y asimétricas podemos saber que clase de imputación vamos a aplicar.



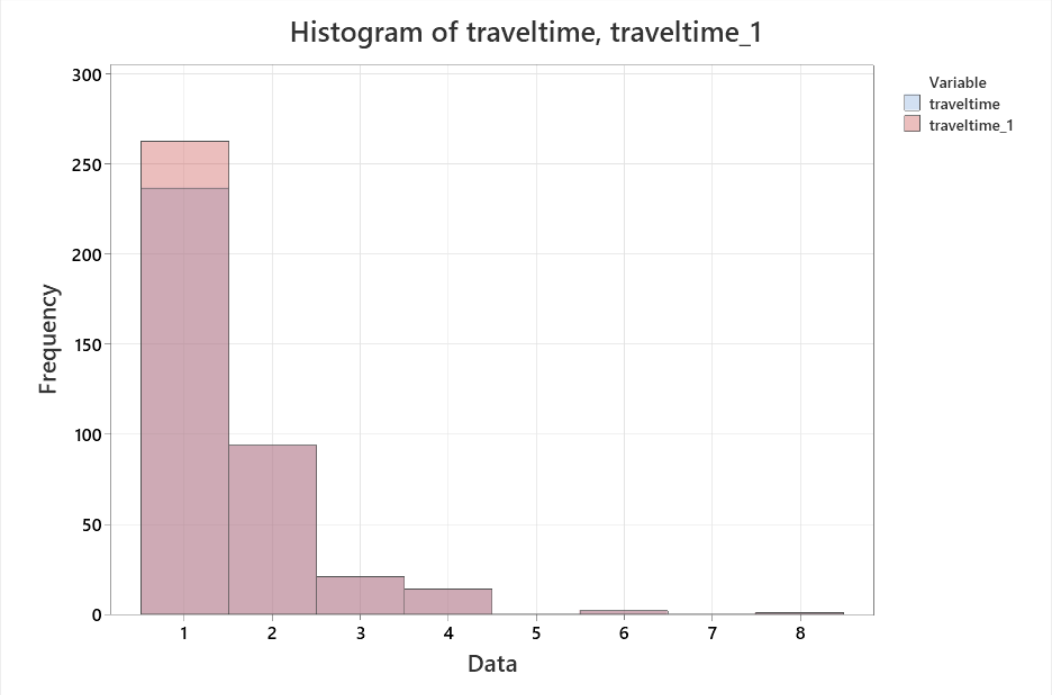
Tomando esto en cuenta sustituimos todos los valores faltantes por la mediana de las respectivas variables.

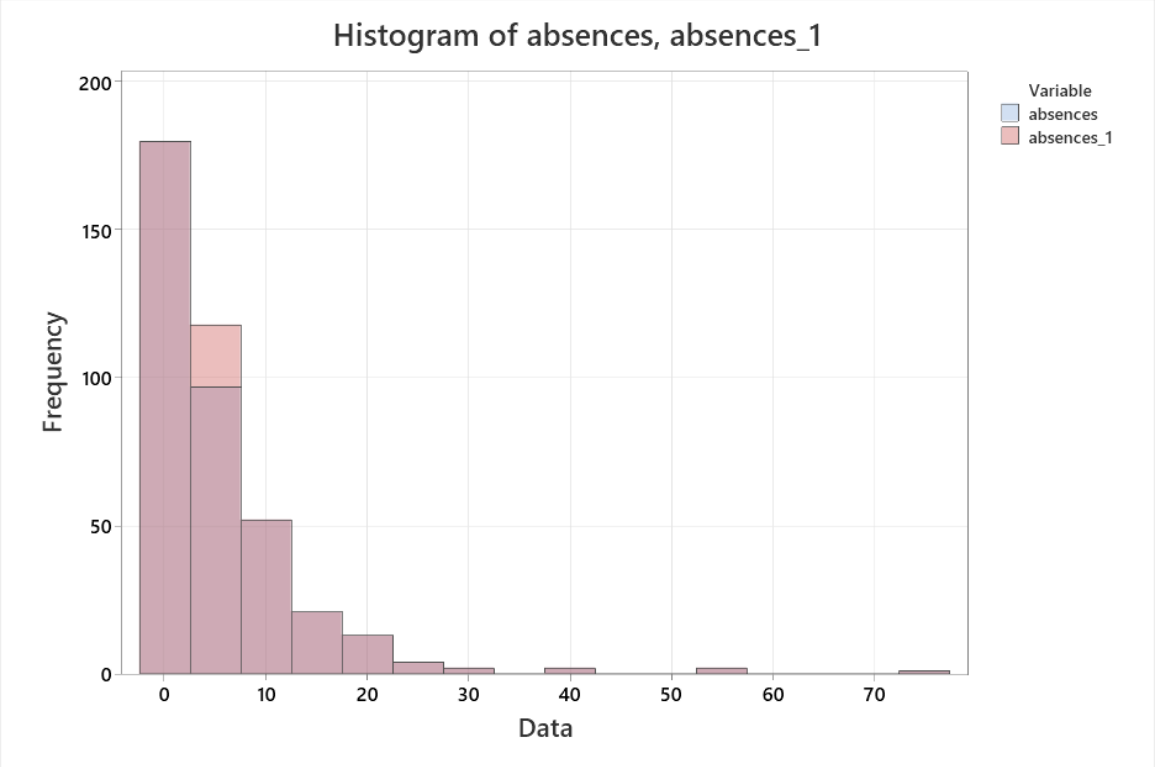
Statistics

Variable	Total Count	N	N*	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
age	395	395	0	16.696	1.276	15.000	16.000	17.000	18.000	22.000
Medu	395	395	0	2.7494	1.0947	0.0000	2.0000	3.0000	4.0000	4.0000
Fedu	395	363	32	2.5207	1.1007	0.0000	2.0000	2.0000	3.0000	4.0000
traveltime	395	369	26	1.5285	0.9028	1.0000	1.0000	1.0000	2.0000	8.0000
studytime	395	395	0	2.1595	1.2594	1.0000	1.0000	2.0000	2.0000	12.0000
failures	395	395	0	0.3342	0.7437	0.0000	0.0000	0.0000	0.0000	3.0000
famrel	395	395	0	3.9443	0.8967	1.0000	4.0000	4.0000	5.0000	5.0000
freetime	395	395	0	3.2354	0.9989	1.0000	3.0000	3.0000	4.0000	5.0000
goout	395	395	0	3.1089	1.1133	1.0000	2.0000	3.0000	4.0000	5.0000
Dalc	395	324	71	1.3580	0.8034	1.0000	1.0000	1.0000	1.0000	5.0000
Walc	395	395	0	2.2911	1.2879	1.0000	1.0000	2.0000	3.0000	5.0000
health	395	395	0	3.5544	1.3903	1.0000	3.0000	4.0000	5.0000	5.0000
absences	395	374	21	5.543	8.089	0.000	0.000	3.500	8.000	75.000

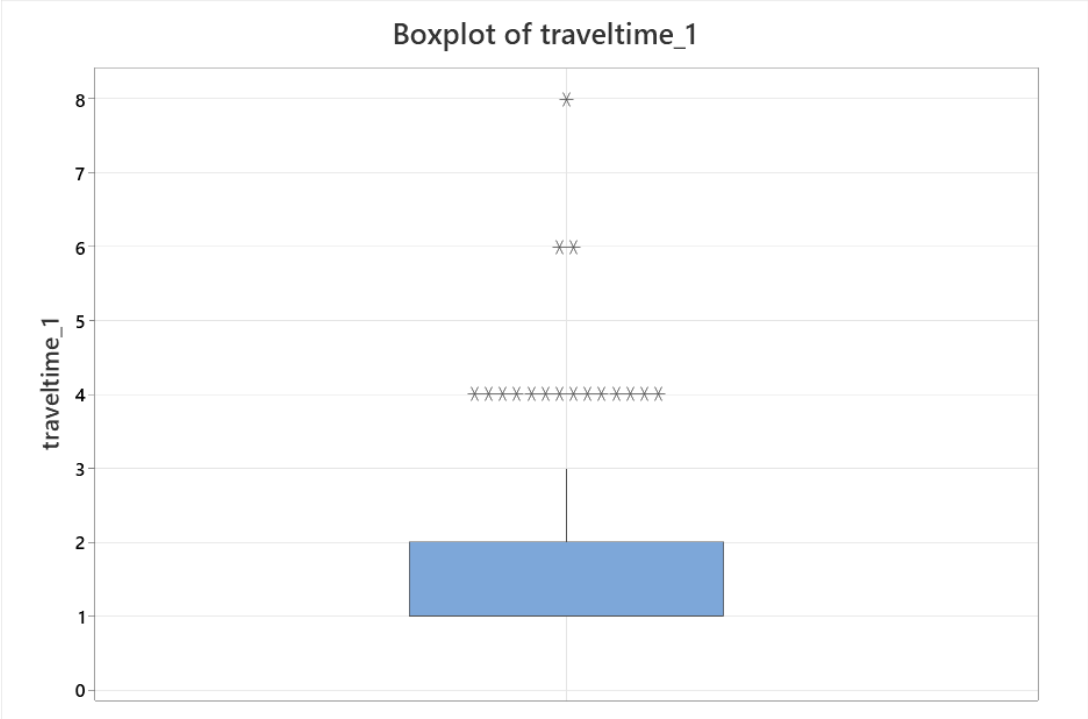
La mediana de traveltime es 1 entonces sustituimos 1 en todos los valores faltantes. La mediana de la variable absences es 3.5, en este caso al 3.5 no ser un valor posible en la variable absences decidimos redondear a 3 ya que la mayoría de los valores de esta variable se encuentran más hacía la izquierda en 0 y 1.

Estos son los histogramas de la comparación entre las variables normales y las variables después de realizar la imputación.

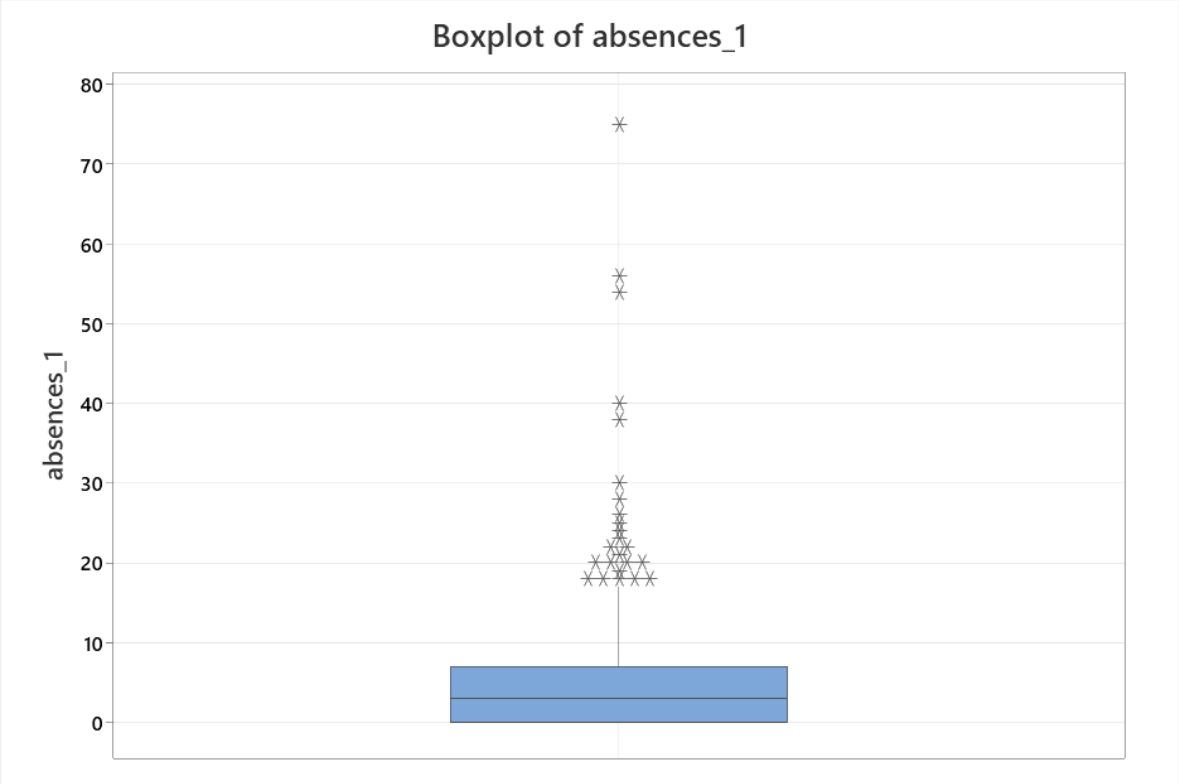




5.Realizar un boxplot e interpretarlo.



Para la variable traveltime tenemos 4 valores posibles: 1, 2, 3 y 4, los cuales representan distintos tiempos en los que los alumnos llegan a la escuela. Como podemos observar en el gráfico de caja la mayoría de los valores se encuentran entre el 1 y el 2, pero podemos ver que tenemos valores **outliers** ya que podemos ver que 14 personas eligieron la opción 4. También podemos ver que hay valores en 6 y 8 pero como estos no son valores válidos simplemente se descartarían.



Por último, tenemos el gráfico de caja de la variable absences, en esta se nos muestra que la mayoría de los valores se encuentran entre 0 y 7 aproximadamente, ya que la caja nos indica en donde inicia el primer cuartil, a la mitad el segundo cuartil y el final de la caja es el tercer cuartil. También podemos ver que tenemos valores outliers, que van desde 17 aproximadamente hasta 75. En este caso el número máximo de faltas que puede tener un alumno es 93 entonces todos nuestros valores outliers son válidos.