Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Guadalajara

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)

M1.2 Datos Faltantes y Outliers

Samuel Padilla Esqueda | A01641383

Agosto 2024

Utiliza la variable absences y traveltime del archivo: student-mat_v3.csv para realizar lo siguiente:

1. **Identificar el porcentaje de datos faltantes.**
   Usando la siguiente formula en Excel
   - =CONTAR.SI(Rango,"NA")/CONTAR(Rango)*100

   Donde Rango son las celdas con los valores de las variables deseadas, obtenemos los porcentajes de datos faltantes al dividir la cantidad de valores "NA" y dividirlo entre el total de valores.

   Traveltime: 7.05%
   Absences: 5.61%

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

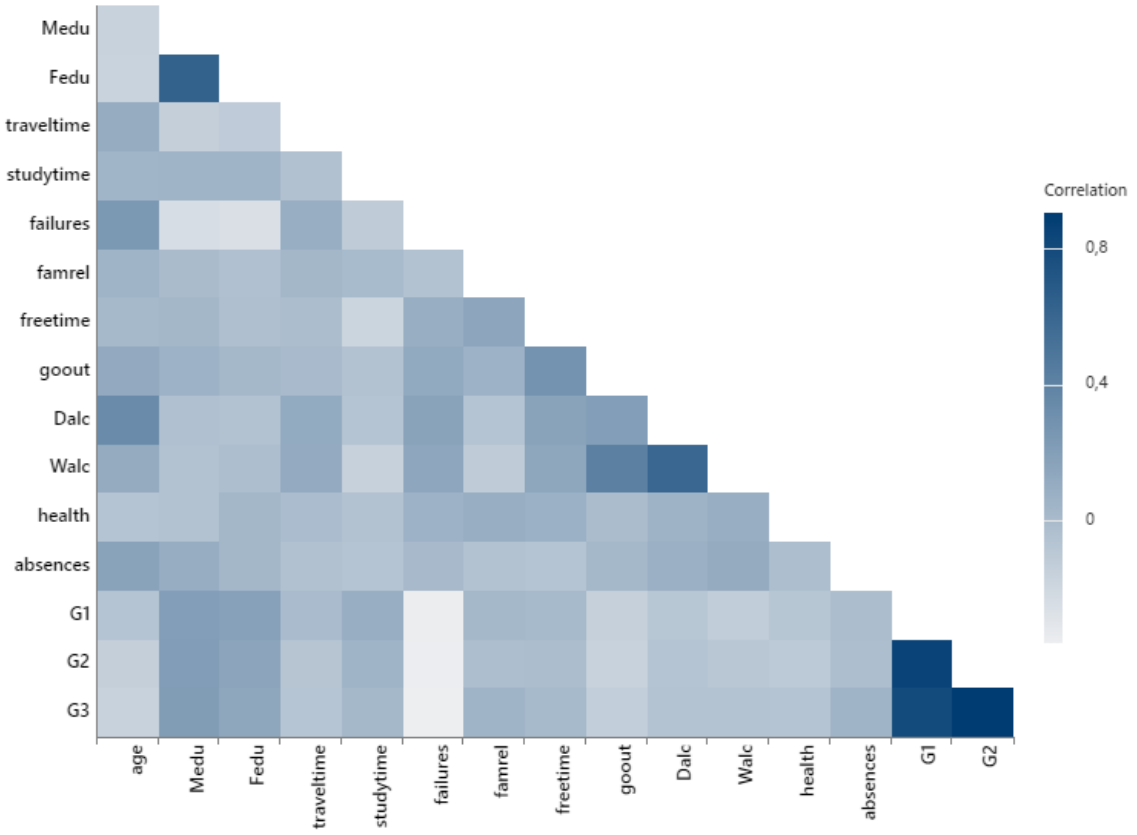
   Para identificar el mecanismo primero calculamos las correlaciones entre traveltime y absences con el resto de las variables, hacemos esto usando minitap:

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
G1	-0,064	0,205	0,190	0,002	0,095	-0,355	0,022	0,013	-0,149
G2	-0,143	0,216	0,160	-0,071	0,054	-0,356	-0,018	-0,014	-0,162
G3	-0,162	0,217	0,147	-0,066	0,024	-0,360	0,051	0,011	-0,133

	Dalc	Walc	health	absences	G1	G2
Medu						
Fedu						
traveltime						
studytime						
failures						
famrel						
freetime						
goout						
Dalc						

Walc	0,598													
health	0,057	0,092												
absences	0,077	0,117	-0,020											
G1	-0,080	-0,126	-0,073	-0,011										
G2	-0,060	-0,085	-0,098	-0,018	0,852									
G3	-0,057	-0,052	-0,061	0,049	0,801	0,905								



Podemos observar que los valores de correlación de **traveltime** con el resto de las variables de muy pequeño, su correlación mas grande es 0,112 con la variable age, pero es un valor tan cercano a 0 que implica una correlación muy débil. Ya que no depende de otra variable, descartamos el mecanismo MAR, por lo que puede ser que aplique el modelo MCAR o NMAR.

La variable **Absences** tiene un caso parecido, donde la mayor correlación se observa con la variable age, con un valor de 0,173. Aunque es un valor muy cercano a 0, también es significativamente mayor que el resto de correlaciones por lo que creo que si puede haber una correlación, aunque no sea muy fuerte, de esto modo descartamos el método MAR, ya que hay evidencia de que Absences depende de otra variable, es posible que se trate de un mecanismo MCAR o MAR.

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

   Usando las fórmulas de Excel podemos obtener estas estadísticas (sin imputación):

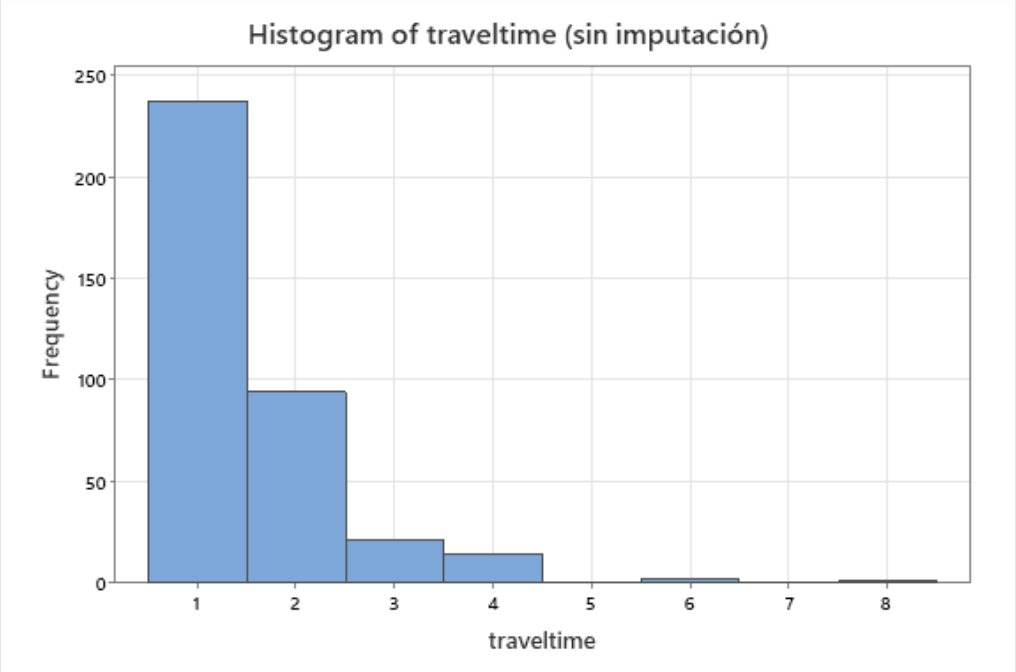
   Traveltime:

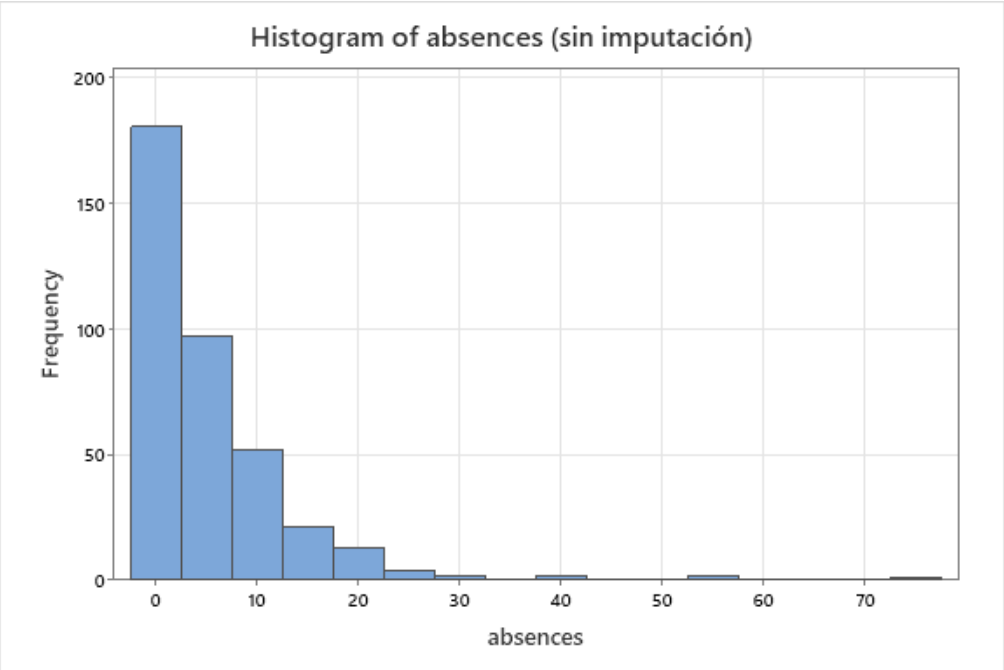
%datos faltantes	7.05
Mec datos faltantes	MCAR o NMAR
Media	1.528455285
desviación estándar	0.9028204892
mediana	1
moda	1

   Absences:

%datos faltantes	5.61
Mec datos faltantes	MCAR o MAR
Media	5.542780749
desviación estándar	8.089117488
mediana	3.5
moda	0

   Usando minitap obtenemos los siguientes histogramas:





4. **Utilizar el método de imputación adecuado para cada una de las variables con datos faltantes.**
    - **Imputación Simple: Media, Mediana, Moda**

Implementamos imputación simple con moda, para poner los datos faltantes en el mismo lugar donde están la mayoría de los datos y evitar que se altere demasiado la distribución de los datos.

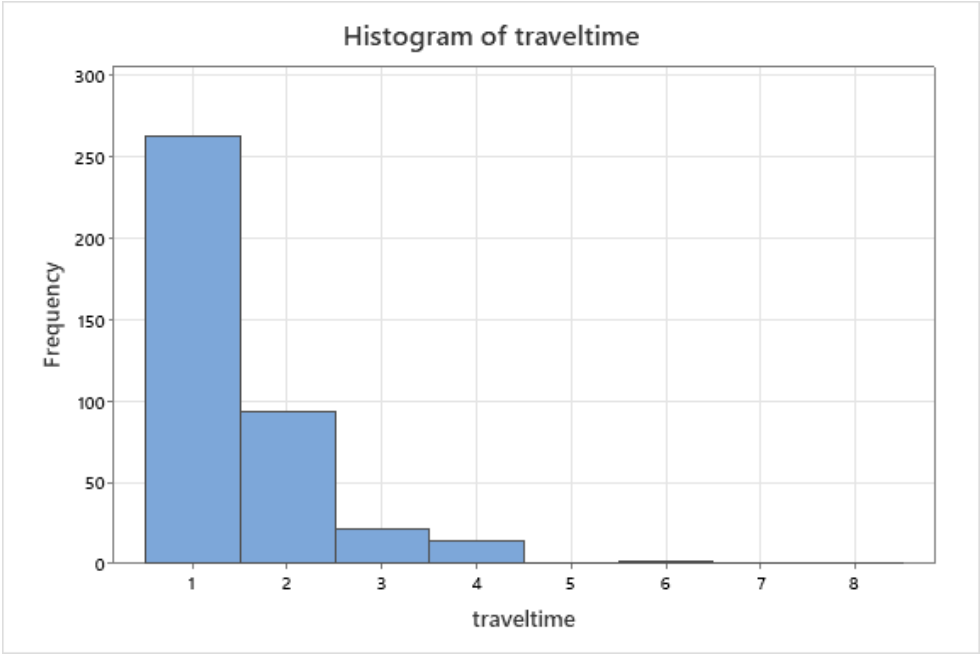
Reemplazamos en Excel los valores NA con el valor de la moda de cada respectiva variable. Obtenemos estas nuevas estadísticas.

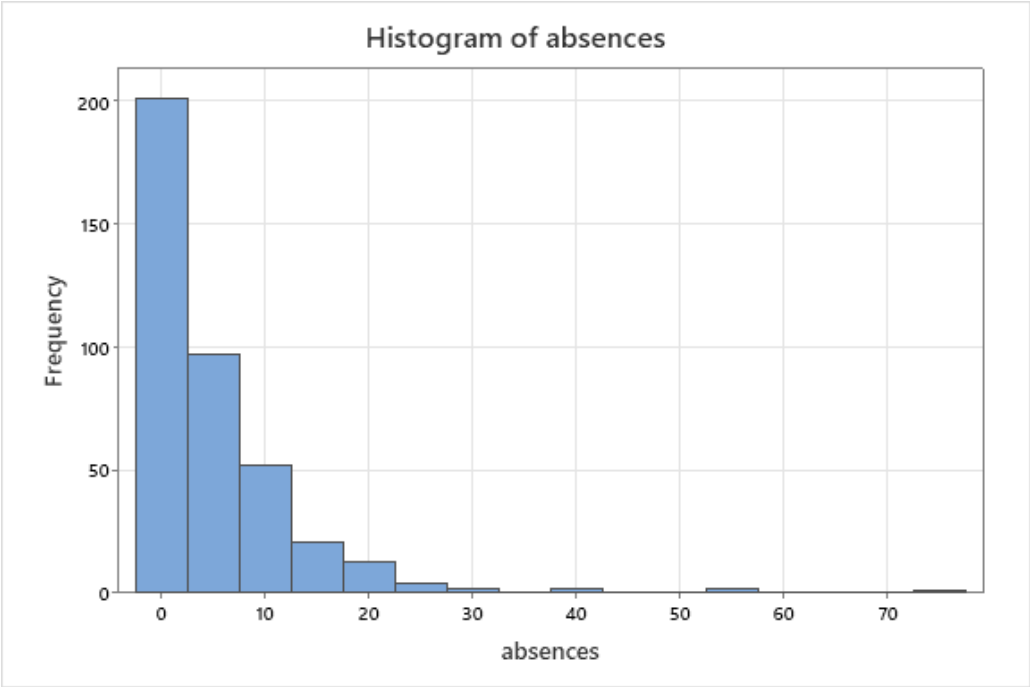
Traveltime:

%datos faltantes	0
Mec datos faltantes	MCAR o NMAR
Media	1.493670886
desviación estándar	0.8823339165
mediana	1
moda	1

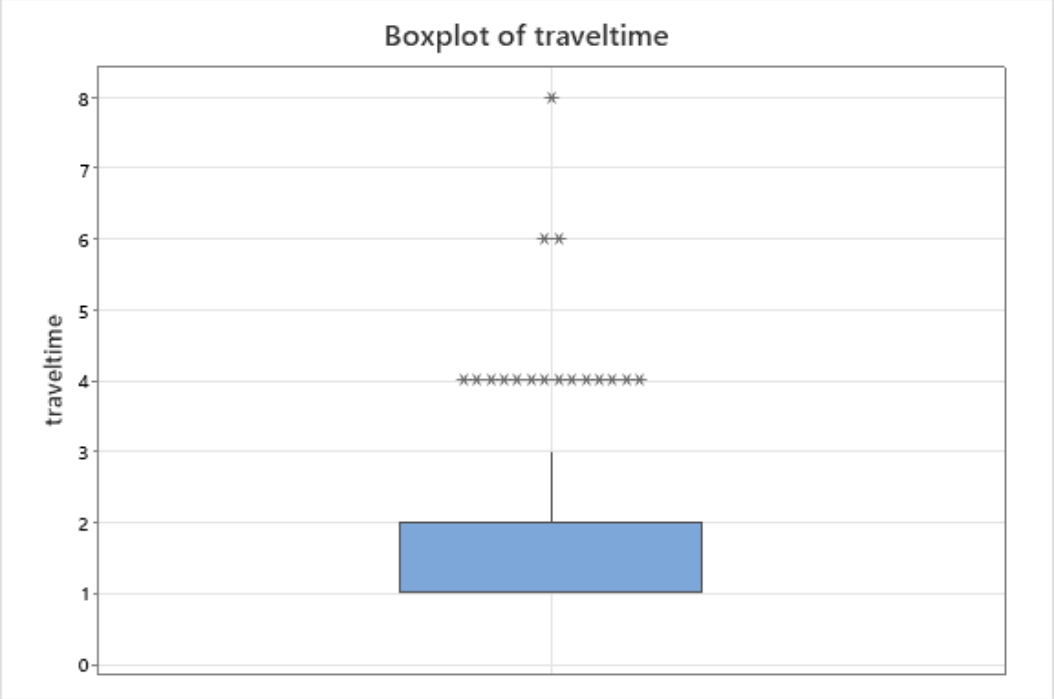
Absences:

%datos faltantes	0
Mec datos faltantes	MCAR o MAR
Media	5.248101266
desviación estándar	7.968479476
mediana	2
moda	0

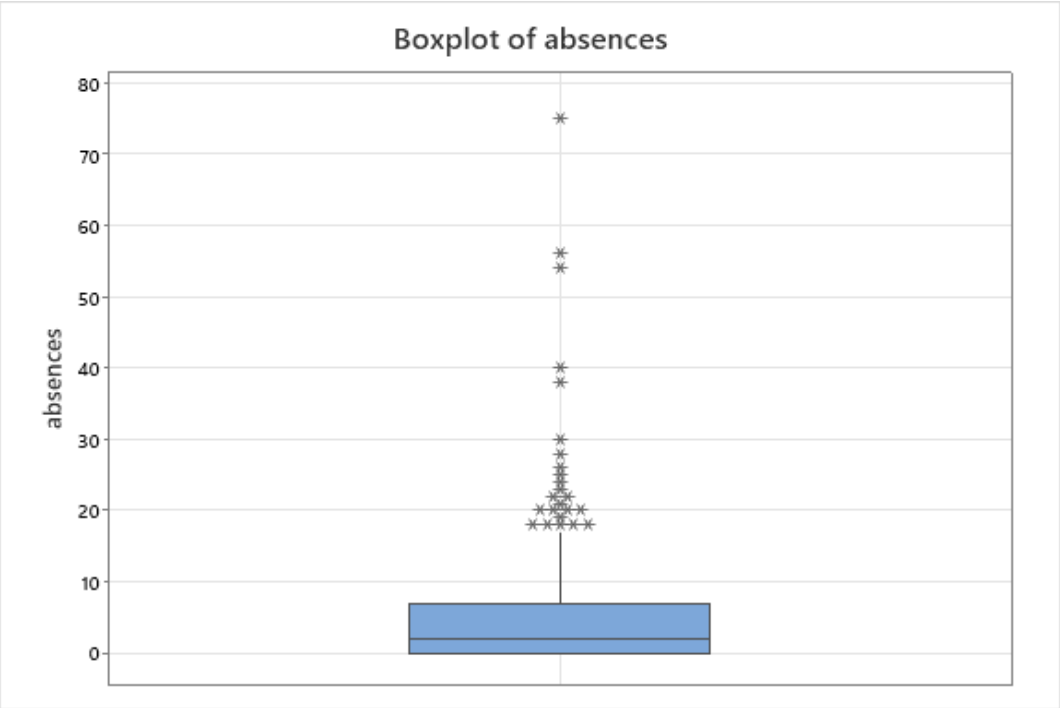

Histogram of traveltime
Frequency
300
250
200
150
100
50
0
1
2
3
4
5
6
7
8
traveltime


Histogram of absences
Frequency
200
150
100
50
0
0
10
20
30
40
50
60
70
absences

5. **Realizar un boxplot e interpretarlo.**
   Usamos boxplot de los datos tras la imputación simple usando la moda.
   Usamos los boxplots para identificar valores atípicos.



Observamos que el 50% de los datos están entre 1 y 2, dentro de la caja.
Hay una gran cantidad de valores atípicos que superan por mucho el promedio de los valores, la mayoría con un valor de 4.



Observamos que el 50% de los datos están entre 0 y 8-9, dentro de la caja.
Hay una gran cantidad de valores atípicos que superan por mucho el promedio de los valores, llegando hasta a mas de 50. Observamos que la mediana se encuentra en la parte inferior de la caja.