Tecnológico de Monterrey

M1.2 Datos Faltantes y Outliers

Roberto Angel Rillo Calva A01642022

Jacob Valdenegro Monzón A01640992

Tecnológico de Monterrey

Inteligencia artificial avanzada para la ciencia de datos I

(Gpo 101)

15 de agosto del 2024

Absences y Traveltime

1.Identificar el porcentaje de datos faltantes.

```
def contar_nulos_columna(df, columna):
 nulos = df[columna].isnull().sum()
 return nulos


columna = 'absences'
nulos = contar_nulos_columna(datos, columna)
porcentaje_nulos = (nulos / len(datos)) * 100
print(f"Cantidad de valores nulos en '{columna}': {nulos}")
print(f"El porcentaje de nulos es del '{porcentaje_nulos}'")
```

Cantidad de valores nulos en **'absences'**: 21
El porcentaje de nulos es del '5.3164556962025316'

Cantidad de valores nulos en **'traveltime'**: 26
El porcentaje de nulos es del '6.582278481012659'

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

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

Al analizar los datos recopilados por travel time y absences y comparados con algunos datos numéricos en el data frame en una tabla de correlación podemos deducir que las correlaciones son en su mayoría bajas o débiles. Por lo tanto, el mecanismo que ocasiona los datos faltantes en estas variables es más probable que sea MCAR (Missing Completely at Random).

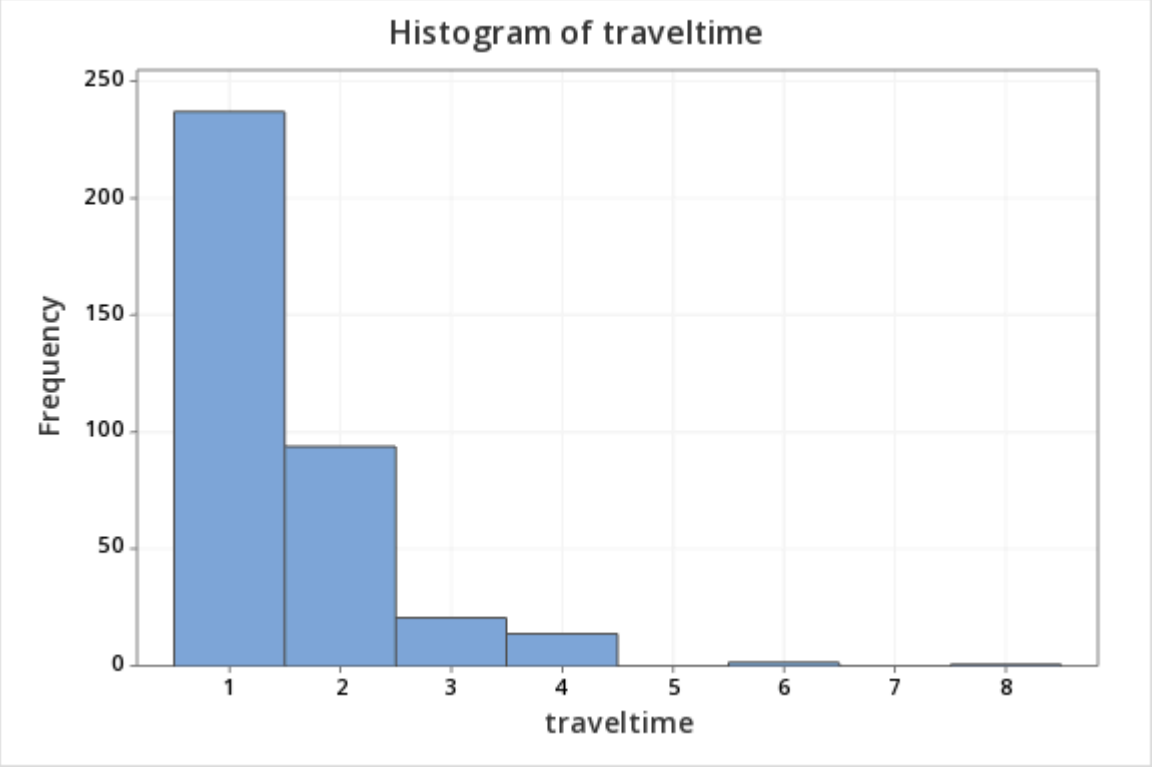
3.Obtener estadísticas descriptivas de los datos (histograma, media, desviación estándar, mediana, moda, etc).

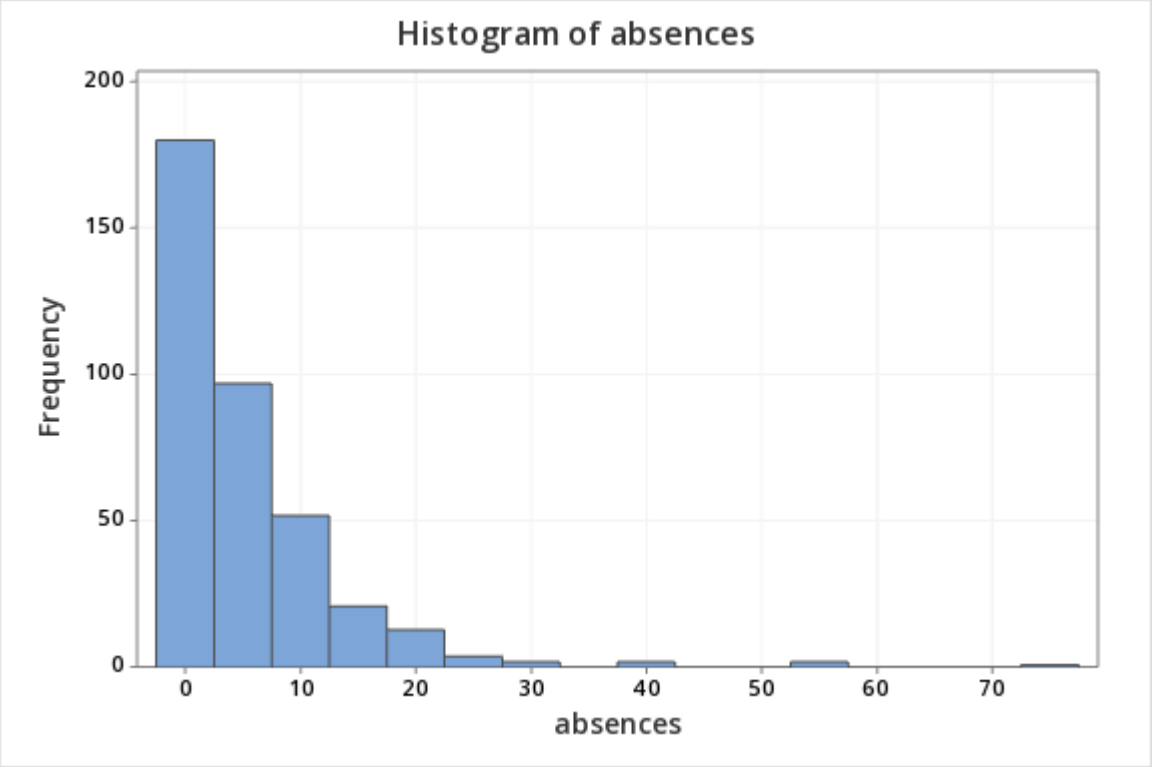
Statistics

Variable	Mean	SE Mean	StDev	Median	Mode	N for Mode
traveltime	1.5285	0.0470	0.9028	1.0000	1	237
absences	5.543	0.418	8.089	3.500	0	115

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

Después del análisis realizado con un histograma, podemos observar que tanto la columna absences como la de travel time tienen una distribución asimétrica, como cada una contiene datos numéricos, se optó por una imputación a los datos faltantes con la **mediana**.





5.Realizar un boxplot e interpretarlo

La mayoría de los estudiantes tiene pocas ausencias, pero hay algunos con un número significativamente alto, lo que podría indicar casos específicos de inasistencia crónica. En cuanto al tiempo de viaje, la mayoría vive cerca de la escuela, pero hay algunos que tienen trayectos mucho más largos, lo que podría influir en su rendimiento y asistencia. Estas observaciones sugieren variabilidad en los comportamientos de asistencia y acceso a la escuela.


Boxplot of Absences
Number of Absences
70
60
50
40
30
20
10
0
1


Boxplot of Absences
Number of Absences
70
60
50
40
30
20
10
0
1