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 García Berenfeld | A01642317
Israel Vidal Paredes | A01750543

13/08/2024

INDEX

PATRONES

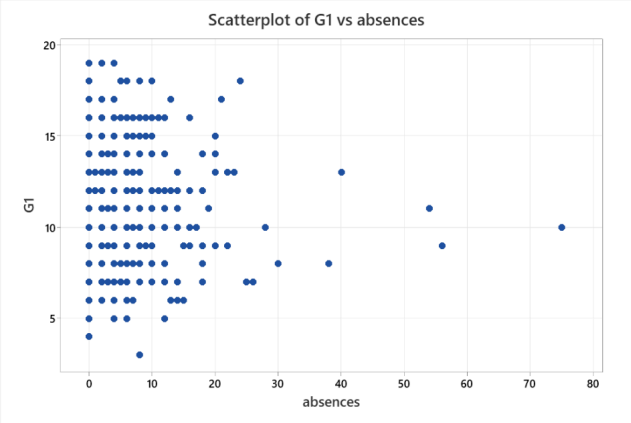
1. Porcentaje de datos faltantes

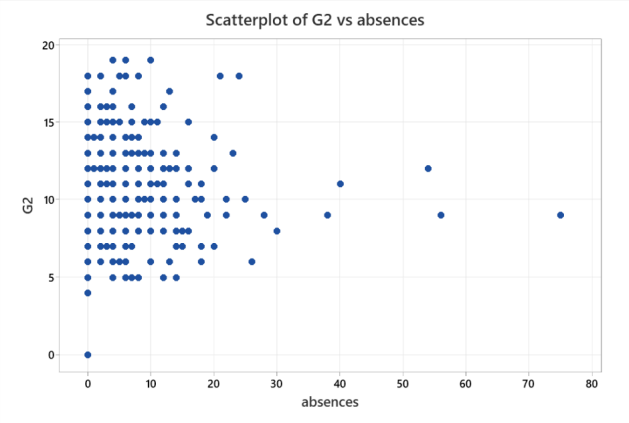
- Absences NA percent: 5.61%
- Travel Time NA percent: 7.05%

2. Hipótesis de mecanismo de datos faltantes

Absences:

Para esta variable, al hacer un gráfico de dispersión comparándolo con las demás variables del modelo podemos observar una tendencia general representada por estos dos gráficos:

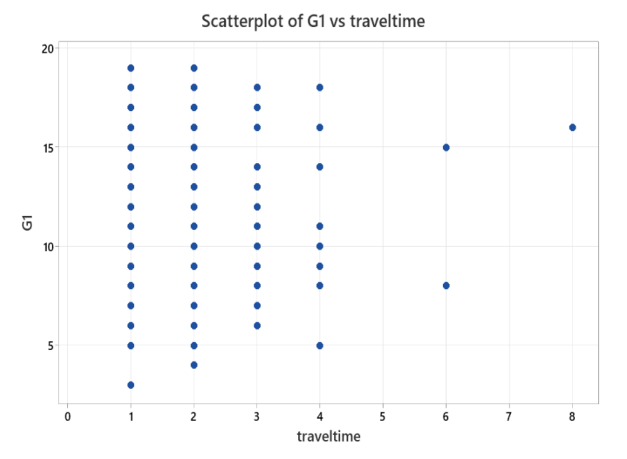


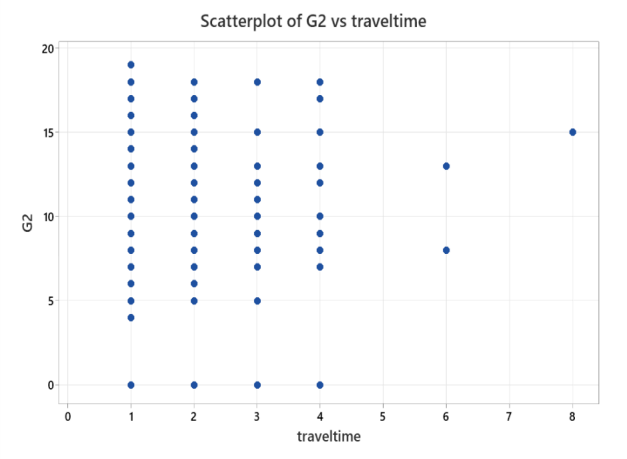


Como podemos observar, los datos se cargan mucho a la izquierda, implicando que algunos de los datos del lado derecho podrían faltar. Este comportamiento nos apunta a MAR como hipótesis del mecanismo que ocasionó la falta de datos. Cabe mencionar que el rango válido de la variable es de 0-93, por lo que los datos del lado derecho no están ahí debido a error.

Travel time:

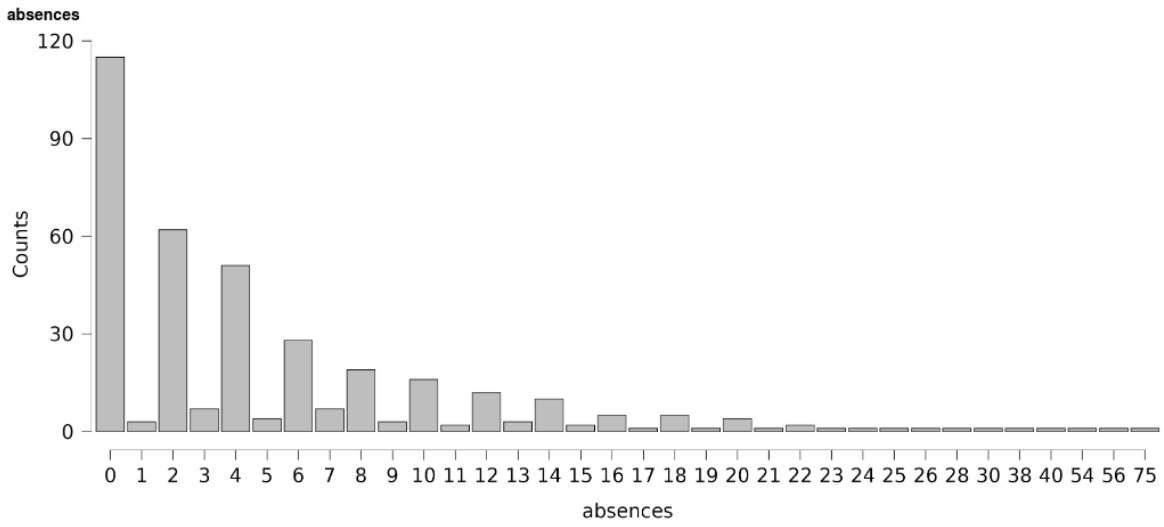
Para este segundo dato, al graficar la variable contra cualquier otra, podemos observar que no encontramos una tendencia general como en el análisis de la variable anterior:





sabemos esto porque, aunque pareciera que también se cargan los datos al lado izquierdo, podemos explicar a los outliers del lado derecho como error, ya que la descripción de los datos nos indica que el valor máximo debería ser 4. Al eliminar datos a la derecha de 4 podemos observar que no hay carga de datos en ninguna dirección, por lo que nuestra hipótesis es que el mecanismo que está dando falta a datos es MCAR.

3. Estadísticas descriptivas


absences
120
90
60
30
0
Counts
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 28 30 38 40 54 56 75
absences

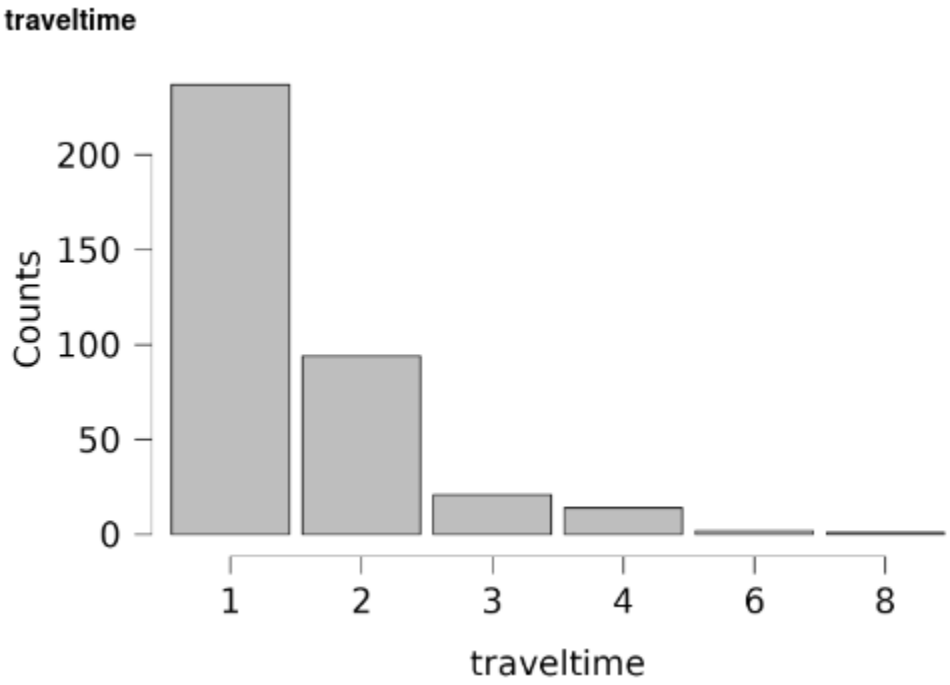

traveltime
200
150
100
50
0
Counts
1
2
3
4
6
8
traveltime

Table 1: Descriptive Statistics

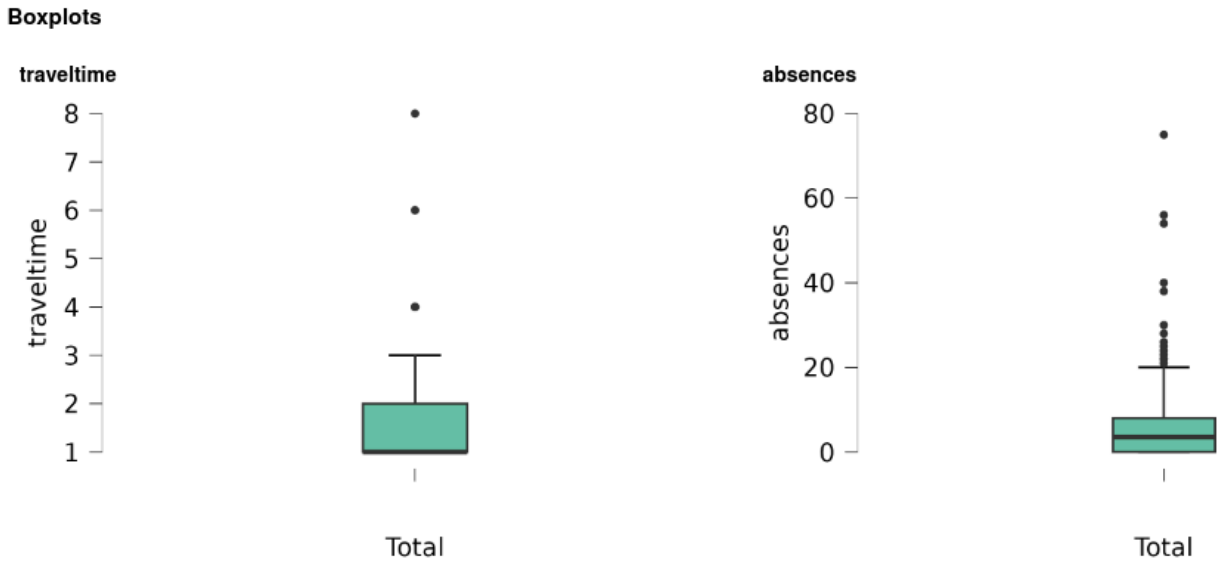
	traveltime	absences
Valid	369	374
Missing	26	21
Mode	1.000	0.000
Median	1.000	3.500
Mean	1.528	5.543
Std. Deviation	0.903	8.089
Coefficient of variation	0.591	1.459
Minimum	1.000	0.000
Maximum	8.000	75.000

4. Método de imputación seleccionado

Media, Mediana, Moda, gráficos comparativos

Debido a que ***absences*** es una variable ordinal que muestra una distribución asimétrica, el método de imputación seleccionado será usando la mediana del conjunto de datos. Mientras que para la variable ***traveltime***, al ser una variable categórica, se usará la moda

5. Boxplots



Representando los datos de ***traveltime*** y ***absences*** en gráficos de cajas, se pueden observar valores atípicos en ambas distribuciones, de los cuales, los valores atípicos mayores a 4 en ***traveltime*** se identifican como erróneos, ya que la variable categórica solamente acepta valores de 1 a 4.