DETECCIÓN DE OUTLIERS

JORGE EDUARDO ESTRADA DAVILA 1741943
ALEJANDRO URIEL GARCÍA ALANÍS 1886968
NÉSTOR MISAEL PAZ REYES 1559508
ELISA GONZÁLEZ GARCÍA 1858207
RENE SOBREVILLA RUIZ 1941452

¿QUÉ SON LOS OUTLIERS?

- Es una observación anormal en una muestra estadística o serie temporal que afecta la estimación de parámetros.

OUTLIERS

- Identificar "outliers" es fundamental si se desea realizar un análisis descriptivo, ya que estos forman parte de la estadística descriptiva.
- Los valores extremos se denominan "**outliers**" en inglés.
- Los valores internos se denominan "**insiders**" en inglés.

SIGNIFICADO DE LOS OUTLIERS

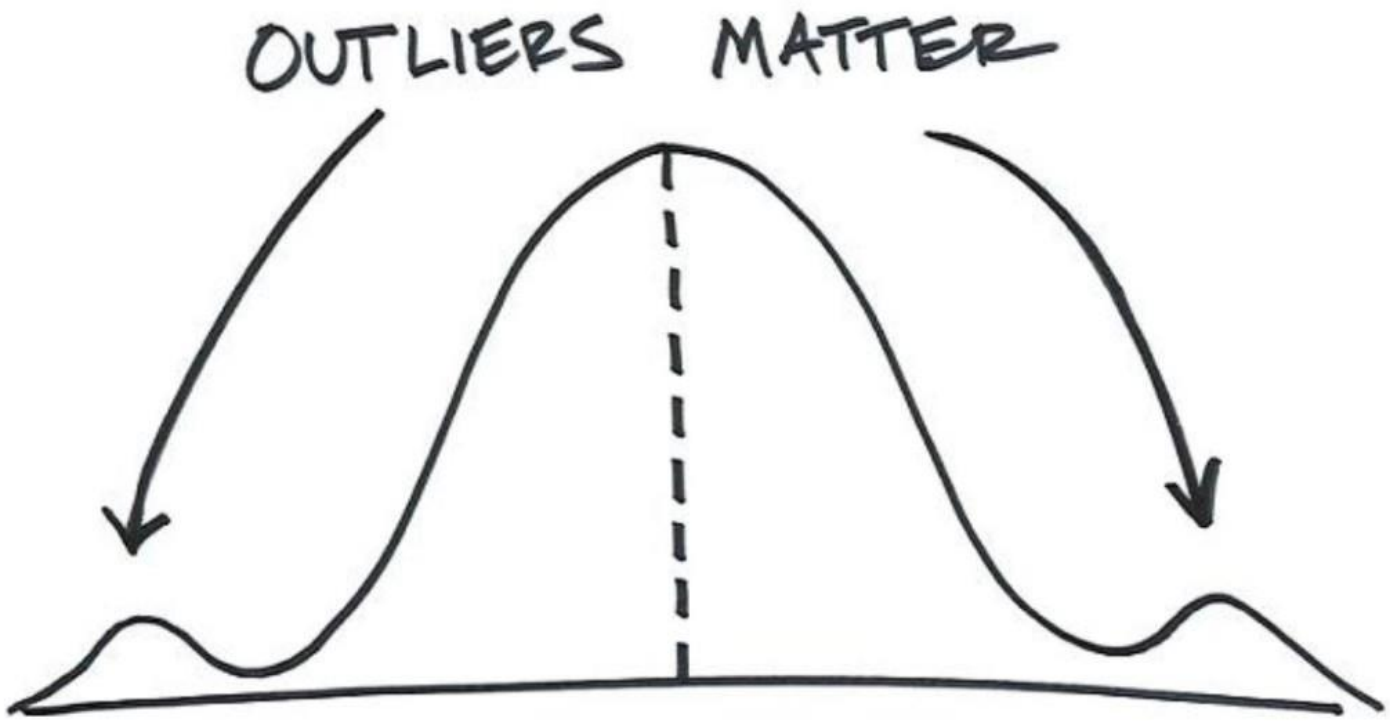
Tener este tipo de valores extremos puede significar:

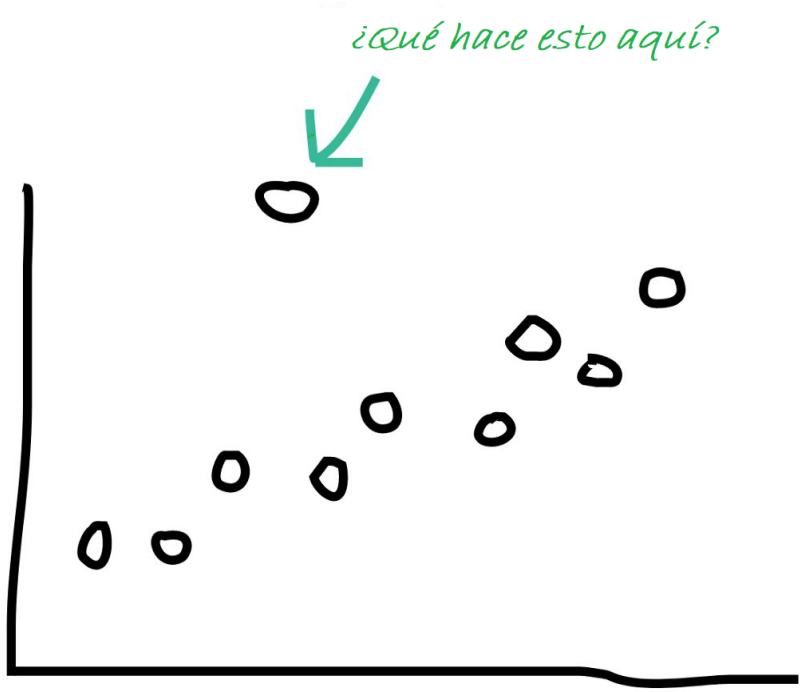
1. **ERROR**: Errores a la hora de capturar datos
2. **CASOS EXTREMOS**: Momentos o casos fuera de lo esperado, casos anómalos

¿CÓMO DETECTAR LOS OUTLIERS?

Hay varias maneras de detectar outliers:

- Mediante hipótesis: Q-test o Grubss-test
- Mediante gráficas: Uso del Boxplox o Histogramas
- Mediante indicadores: Usando la normal estándar o desviación estándar de los datos





MÉTODOS DE DETERMINACIÓN DE OUTLIERS

N	Q_{crit} (CL:90%)	Q_{crit} (CL:95%)	Q_{crit} (CL:99%)
3	0.941	0.970	0.994
4	0.765	0.829	0.926
5	0.642	0.710	0.821
6	0.560	0.625	0.740
7	0.507	0.568	0.680
8	0.468	0.526	0.634
9	0.437	0.493	0.598
10	0.412	0.466	0.568

$$Q_{TS} = \frac{\text{Gap}}{\text{Range}}$$

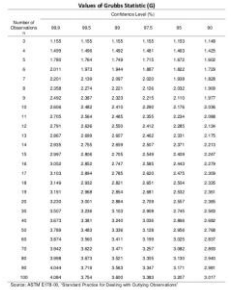
$\mathbf{H}_0$ There are no outliers present in the data.

$\mathbf{H}_1$ There is one outlier present

Q-TEST

- Se ordenan los datos de forma ascendente y, creando un estadístico para cada dato, se hace una prueba de hipótesis si ese dato es o no un "outliers".

$$G = \frac{|Questionable\ data - \bar{x}|}{s}$$



If $G_{calc} > G_{table}$ you can throw out the data point.

GRUBBS-TEST/CRITERIO DE CHAUVENET

- Este es un test que determina si hay o no datos "outliers", construyendo un estadístico para cada dato y compararlo con tablas.

MEDIANTE LA DISTRIBUCIÓN NORMAL

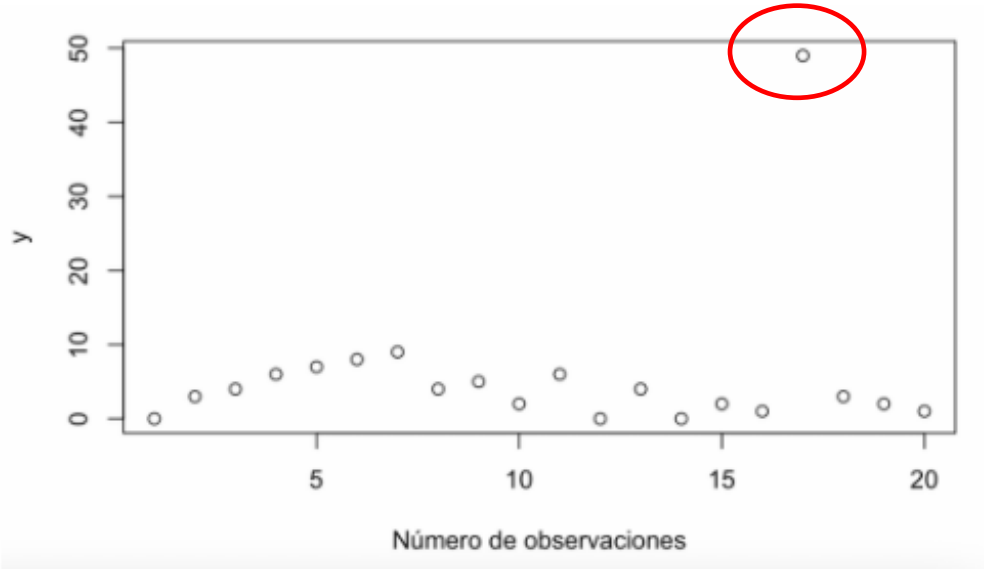
- Podemos estandarizar nuestros datos, caulculando tanto la media como desviación estándar de estos. Con esto, tomamos una región de rechazo con un Alpha deseable para nosotros. Por lo general, se usa que estos valores estandarizados estén entre $-3 \leq Z \leq 3$.

$$Z = \frac{x - \mu}{\sigma}$$

EJEMPLO

Supongamos tenemos los siguientes datos:

0 3 4 6 7 8 9 4 5 2 6 0 4 0 2 1 49 3 2 1



- Podemos observar que el valor que está más alejado del resto puede ser muy probablemente un outlier.

EJEMPLO

Usando R, podemos realizar lo siguiente

- Primero calculamos la media y la desviación típica:
- x = media = 5,8
- sigma = desviación típica = 10,51
- Luego sustituimos los valores en la fórmula y calculamos el valor de z para cada observación:

Forma para estandarizar los datos:

$$z = \frac{y - x}{\sigma}$$

```
-0.55179516 -0.26638387 -0.17124677  0.01902742  0.11416452  0.20930161  0.30443871 -0.17124677
-0.07610968 -0.36152097  0.01902742 -0.55179516 -0.17124677 -0.55179516 -0.36152097 -0.45665806
 4.10992256 -0.26638387 -0.36152097 -0.45665806
```

EJEMPLO

Los valores anteriores son los factores multiplicativos de sigma, es decir, z. Cualquiera que sea mayor que 3 o menor que -3 será un valor extremo.

```
 0  3  4  6  7  8  9  4  5  2  6  0  4  0  2  1 49  3  2  1

-0.55179516 -0.26638387 -0.17124677  0.01902742  0.11416452  0.20930161  0.30443871 -0.17124677
-0.07610968 -0.36152097  0.01902742 -0.55179516 -0.17124677 -0.55179516 -0.36152097 -0.45665806
 4.10992256 -0.26638387 -0.36152097 -0.45665806
```

Por tanto, el valor extremo u "outlier" del conjunto de datos sería el 49.

POR DESVIACIÓN ESTÁNDAR

- Se saca la desviación estándar y media de los datos, y se checa que la distancia entre el dato y la media sea menor a la desviación estándar por una "n"

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

EJEMPLO

Artículo	Precio
1	$ 230.00
2	$ 210.00
3	$ 250.00
4	$ 200.00
5	$1,000.00
6	$ 200.00
7	$ 200.00
8	$1,000.00
9	$ 210.00
10	$ 200.00

$$\bar{X} = \frac{230 + 210 + 250 + \cdots + 1000 + 210 + 200}{10} = 370$$

$$S = \sqrt{\frac{(230 - 370)^2 + \cdots + (200 - 370)^2}{10 - 1}} = 332.43$$

$$Supongamos\ que\ queremos\ una\ n = 1.5$$

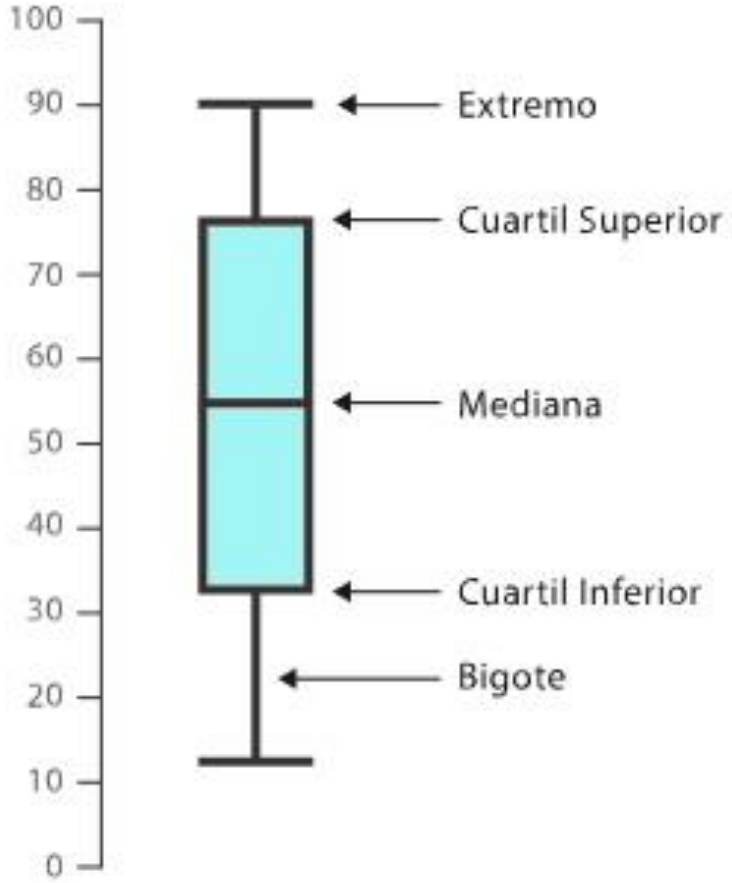
Artículo	Precio	ABS(Precio-Media)	n*S	¿(P-M)>n*S?
1	$ 230.00	140	498.648173	NO
2	$ 210.00	160	498.648173	NO
3	$ 250.00	120	498.648173	NO
4	$ 200.00	170	498.648173	NO
5	$ 1,000.00	630	498.648173	SI
6	$ 200.00	170	498.648173	NO
7	$ 200.00	170	498.648173	NO
8	$ 1,000.00	630	498.648173	SI
9	$ 210.00	160	498.648173	NO
10	$ 200.00	170	498.648173	NO

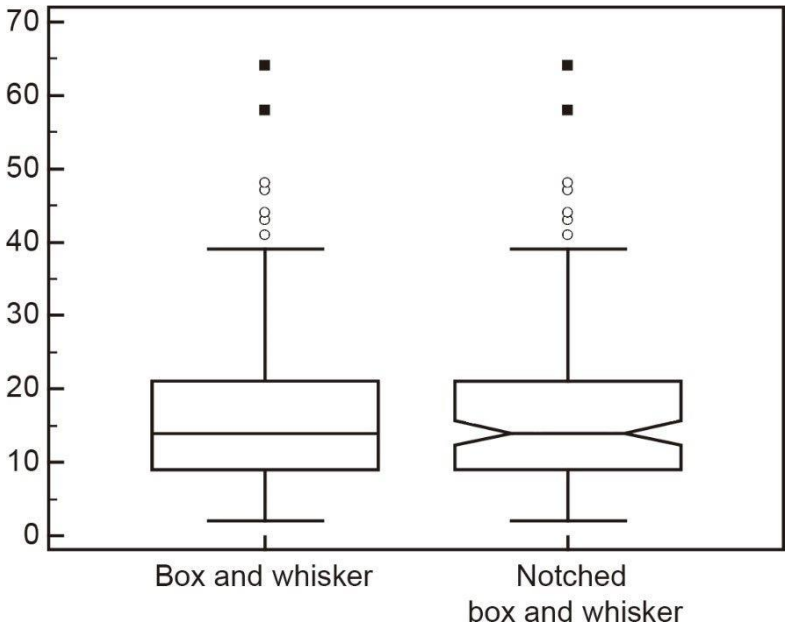
Por lo tanto, el artículo 5 y 8 se consideran "outliers"

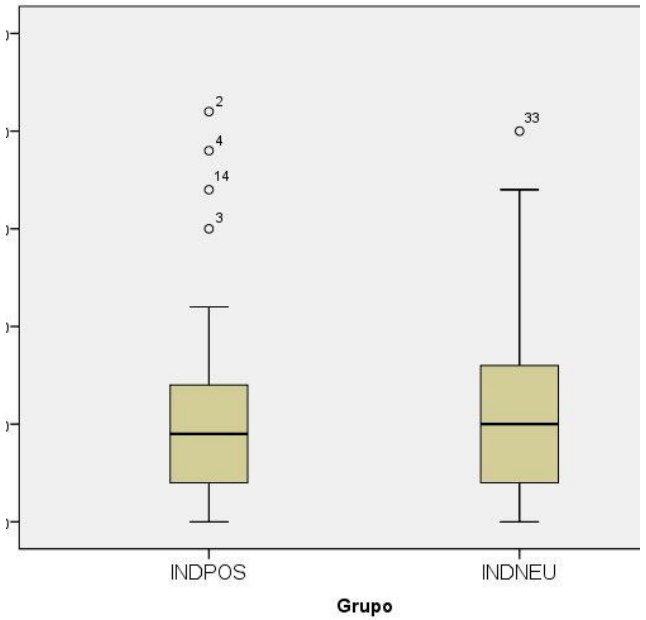
$NOTA: La\ n\ debe\ estar\ entre\ 1 \leq n \leq 3, donde\ 1\ es\ más\ conservador\ y\ 3\ más\ liberal$

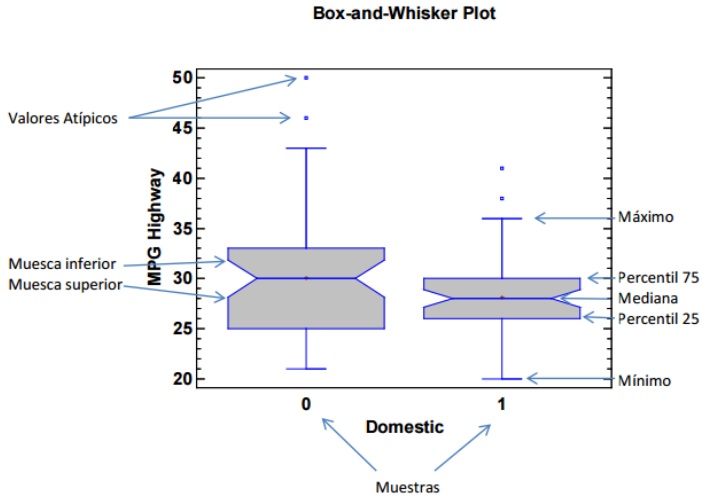
MÉTODO GRÁFICO

- Se usa el "boxplot" o diagrama de caja y bigotes para identificar si existen o no datos "outliers"




70
60
50
40
30
20
10
0
Box and whisker
Notched
box and whisker


2
4
14
3
33
INDPOS
INDNEU
Grupo


Box-and-Whisker Plot
Valores Atípicos
Muesca inferior
Muesca superior
MPG Highway
50
45
40
35
30
25
20
Máximo
Percentil 75
Mediana
Percentil 25
Mínimo
0
1
Domestic
Muestras

EJEMPLOS

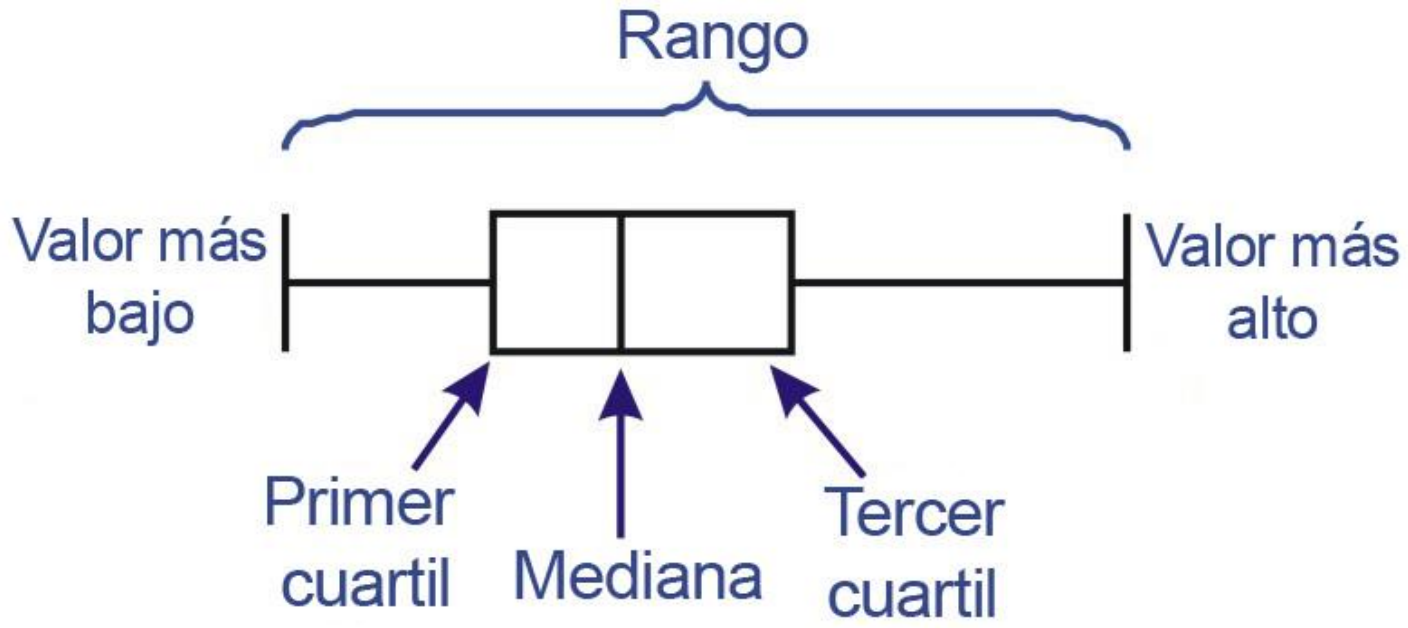
CONSTRUCCIÓN DE BIGOTES DEL "BOXPLOT"

- $Valor\ más\ bajo = Q1 - 1.5(IQR)$
- $Valor\ más\ alto = Q3 + 1.5(IQR)$
- $IQR = Rango\ Intercuartil = Q3 - Q1$

NOTAS:

Si el MIN>Valor mas bajo, entonces el valor mas bajo es el MIN

Si el MAX<Valor mas alto, entonces el valor mas alto es el MAX



EJEMPLO

Librerías usadas

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

Datos usados

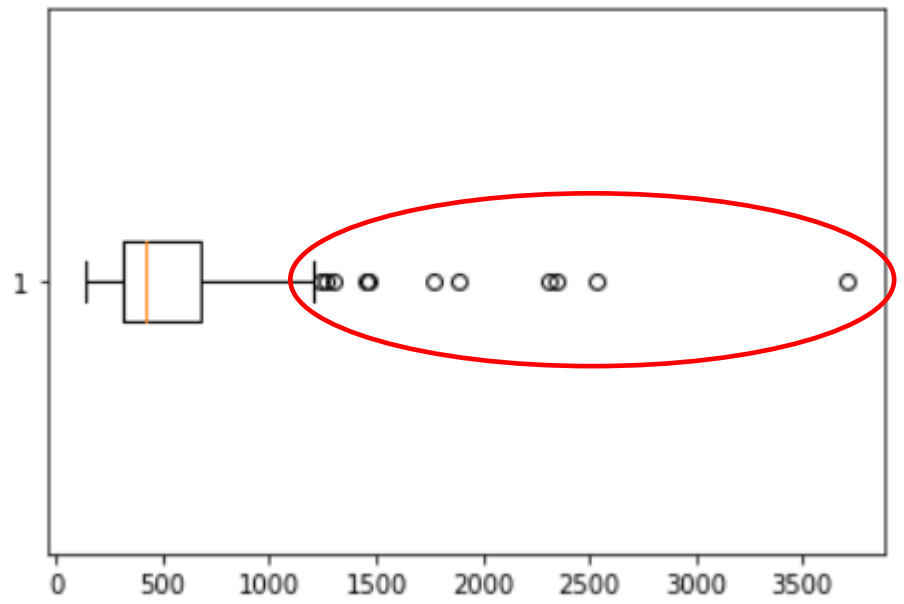
```
Rios = pd.read_csv("RIOS.csv", index_col=0)
Rios
```

Rio	Millas
1	735
2	320
3	325
4	392
5	524
...	...
137	720
138	270
139	430
140	671
141	1770

141 rows × 1 columns

EJEMPLO

```
plt.boxplot(Rios["Millas"], vert=False)
plt.show()
```



Vemos que hay valores “outliers”

EJEMPLO

```
Q1=Rios["Millas"].quantile(0.25)
print("Primer cuartil: ",Q1)
Q3=Rios["Millas"].quantile(0.75)
print("Tercer cuartil: ",Q3)
IQR=Q3-Q1
print("Rango intercuartil: ",IQR)
Valor_Min=Rios["Millas"].min()
print("Valor mínimo: ",Valor_Min)
BI_Calculado=Q1-1.5*IQR
print("Bigote inferior: " ,BI_Calculado)
Valor_Max=Rios["Millas"].max()
print("Valor máximo: ",Valor_Max)
BS_Calculado=Q3+1.5*IQR
print("Bigote superior: " ,BS_Calculado)
```

```
Primer cuartil:  310.0
Tercer cuartil:  680.0
Rango intercuartil:  370.0
Valor mínimo:  135
Bigote inferior:  -245.0
Valor máximo:  3710
Bigote superior:  1235.0
```

Cálculo de cuartiles y Rango intercuartil

EJEMPLO

```
[5] ubicacion_outliers=(Rios["Millas"]<BI_Calculado) | (Rios["Millas"]>BS_Calculado)
    print("Ubicación de outliers\n", ubicacion_outliers)
```

```
Ubicación de outliers
 Rio
1      False
2      False
3      False
4      False
5      False
       ...
137    False
138    False
139    False
140    False
141     True
Name: Millas, Length: 141, dtype: bool
```

Identificar outliers

EJEMPLO

```
outliers=Rios[ubicacion_outliers]
print("Lista outliers\n", outliers)
```

```
Lista outliers
      Millas
Rio
7       1459
23      1450
25      1243
66      2348
68      3710
69      2315
70      2533
83      1306
98      1270
101     1885
141     1770
```

Visualización de outliers

EJEMPLO

```
ubicacion_sin_outliert= (Rios["Millas"] >= BI_Calculado) & (Rios["Millas"] <= BS_Calculado)
Sin_outliers=Rios[ubicacion_sin_outliert]
Sin_outliers
```

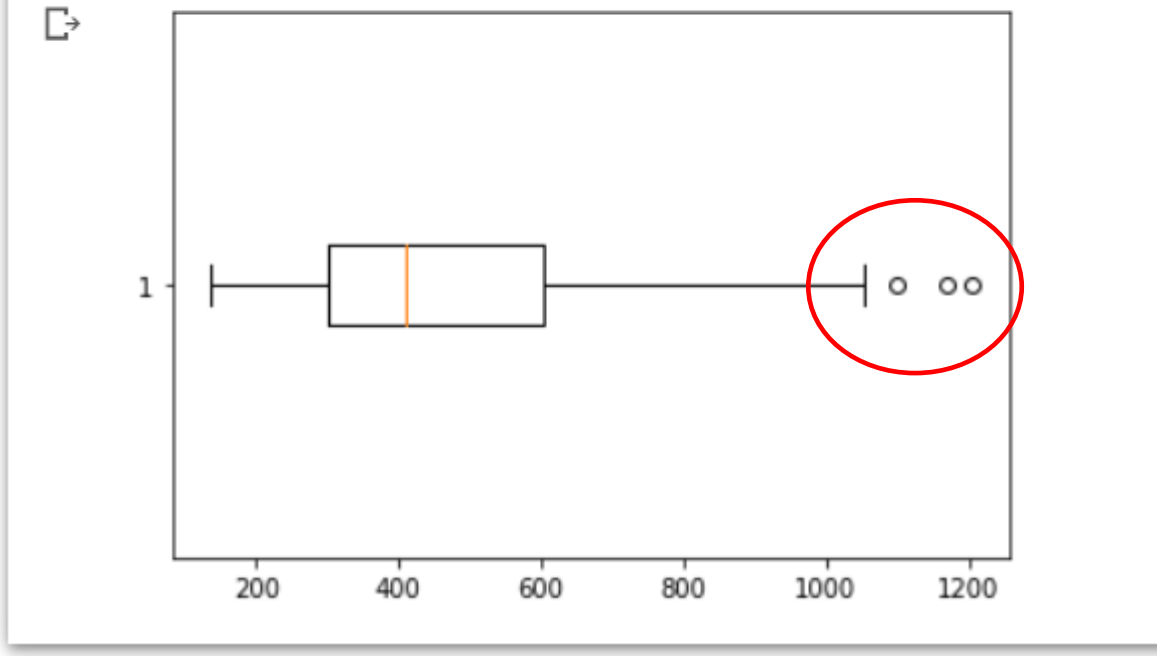
Rio	Millas
1	735
2	320
3	325
4	392
5	524
...	...
136	500
137	720
138	270
139	430
140	671

130 rows × 1 columns

Visualización de datos sin outliers

EJEMPLO

```
plt.boxplot(Sin_outliers["Millas"], vert=False)
plt.show()
```

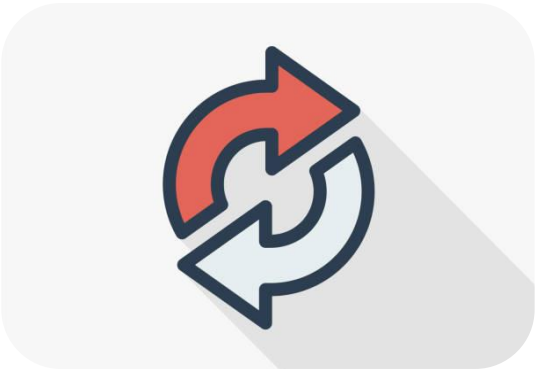


Se debe volver a depurar los outliers

¿QUÉ HACER CON LOS OUTLIERS?

Eliminarlo

Modificarlo



Reemplazarlo

Dejarlo

REFERENCIAS

- Paula Rodó (04 de mayo, 2021).

  Detectar outliers mediante la distribución normal. Economipedia.com
- Francisco Javier Marco Sanjuán (07 de noviembre, 2018).

  Outlier. Economipedia.com
- Na8. (2020). Detección de outliers en Python, de Aprende Maching Learning Sitio web: https://www.aprendemachinelearning.com/deteccion-de-outliers-en-python-anomalia/
- Dr. Masami Yamamoto. (2017). Análisis de los datos: detección de outliers. R E V I S T A E L E C T R Ó N I C A C I E N T Í F I C A Y A C A D É M I C A D E C L Í N I C A A L E M A N A,. Páginas: 31-33.

PREGUNTAS DE INTERÉS

1. ¿Qué son los outliers?
2. ¿Qué tipo de gráficos ayudan a visualizar los outliers?
3. ¿Se pueden detectar outliers en variables categóricas (variables de 0 y 1)?
4. Cuando se detecta un outlier, ¿este debe ser siempre eliminado?
5. ¿Qué es el rango intercuartil?