DETECCIÓN DE OUTLIERS
MINERÍA DE DATOS

- Hector Yair Garza Amaya, 1860264
- Brandon Daniel Alsina Rodriguez, 1860749
- Sebastian Gonzalez Curiel, 1941456
- Alberto Saucedo Orozco, 1735567

¿QUÉ SON LOS OUTLIERS?

- Es interesante ver las traducciones de "outlier" -según su contexto- en inglés:

  - Atípico
  - Destacado
  - Excepcional
  - Anormal
  - Valor Extremo, Valor anómalo, valor aberrante!!

- Es decir, que los outliers en nuestro dataset serán los valores que se "escapan al rango en donde se concentran la mayoría de muestras". En otras palabras, son *las muestras que están distantes de otras observaciones*.

SIGNIFICADO DE OUTLIERS

Los Outliers pueden significar varias cosas:

- **ERROR**: Si tenemos un grupo de "edades de personas" y tenemos una persona con 160 años, seguramente sea un error de carga de datos. En este caso, la detección de outliers nos ayuda a detectar errores.

- **LÍMITES**: En otros casos, podemos tener valores que se escapan del "grupo medio", pero queremos mantener el dato modificado, para que no perjudique la información.

- **Punto de Interés**: puede que sean los casos "anómalos" los que queremos detectar y que sean nuestro objetivo

Muchas veces es sencillo identificar los outliers en gráficas.

DETECCIÓN DE OUTLIERS

Outliers en 1 dimensión

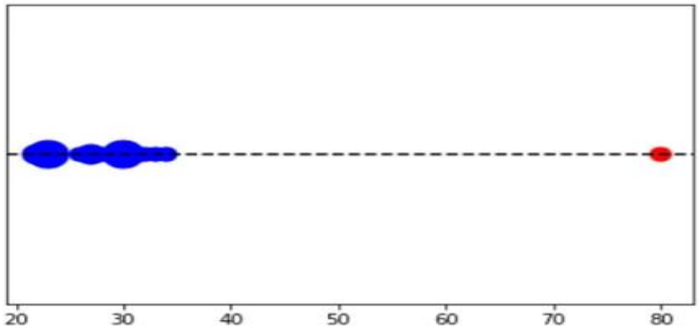
Si analizamos una sola variable, por ejemplo “edad”, veremos donde se concentran la mayoría de muestras y los posibles valores “extremos”.

```python
import matplotlib.pyplot as plt
import numpy as np

edades = np.array([22,22,23,23,23,23,26,27,27,28,30,30,30,30,31,32,33,34,80])
edad_unique, counts = np.unique(edades, return_counts=True)

sizes = counts*100
colors = ['blue']*len(edad_unique)
colors[-1] = 'red'

plt.axhline(1, color='k', linestyle='--')
plt.scatter(edad_unique, np.ones(len(edad_unique)), s=sizes, color=colors)
plt.yticks([])
plt.show()
```



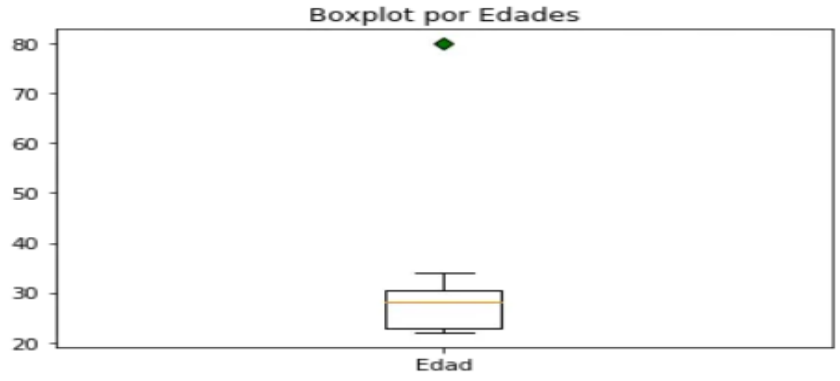
En azul los valores donde se concentra la mayoría de nuestras filas. En rojo un outlier, ó "valor extremo".

En el código, importamos librerías, creamos un array de edades con Numpy y luego contabilizamos las ocurrencias. Al graficar vemos donde se concentran la mayoría de edades, entre 20 y 35 años. Y una muestra aislada con valor 80.

UNA GRÁFICA DE DETECCIÓN SENCILLA: BOXPLOTS

Una gráfica bastante interesante de conocer es la de los Boxplots, muy utilizados en el mundo financiero. En nuestro caso, podemos visualizar las variables y en esa "cajita" veremos donde se concentra el 50 por ciento de nuestra distribución (percentiles 25 a 75), los valores mínimos y máximos (las rayas en "T") y -por supuesto- los outliers, esos "valores extraños" y alejados.

```
1 green_diamond = dict(markerfacecolor='g', marker='D')
2 fig, ax = plt.subplots()
3 ax.set_title('Boxplot por Edades')
4 ax.boxplot(edades, flierprops=green_diamond, labels=["Edad"])
```



¿QUÉ HACER UNA VEZ DETECTADOS?

Según la lógica de negocio podemos actuar de una manera u otra.

Por ejemplo podríamos decidir:

- Las edades fuera de la distribución normal, eliminar.
- El salario que sobrepasa el límite, asignar el valor máximo (media + 2 sigmas).
- Las compras mensuales, mantener sin cambios.

LIBRERÍA PYOD

Una librería muy recomendada es PyOD. Posee diversas estrategias para detectar Outliers. Ofrece distintos algoritmos, entre ellos Knn, el cual analiza la cercanía entre muestras. Veamos cómo utilizarla en nuestro ejemplo.

```
!pip install pyod  # instala la librería

from pyod.models.knn import KNN
import pandas as pd

X = pd.DataFrame(data={'edad':edades,'salario':salario_anual_miles, 'compras':compras_mes})

clf = KNN(contamination=0.18)
clf.fit(X)
y_pred = clf.predict(X)
X[y_pred == 1]
```

	compras	edad	salario
3	20	23	21
10	5	30	56
16	2	33	2
18	2	80	23

LIBRERÍA PYOD

La librería PyOd detecta los registros anómalos.
Para problemas en la vida real, con múltiples dimensiones conviene apoyarnos en una librería como esta que nos facilitará la tarea de detección y limpieza/transformación del dataset.

CONCLUSIONES

Hemos visto lo importante que son los outliers y el impacto que pueden tener en nuestras bases de datos. La mayoría de los datasets tendrán muestras "fuera de rango", por lo que debemos tenerlas en cuenta y decidir cómo tratarlas.

Para algunos problemas, nos interesa detectar esos outliers y de hecho será nuestro objetivo localizar esas anomalías.

Para trabajar con muestras con decenas o cientos de dimensiones nos conviene utilizar una librería como PyOd que realiza muy bien su trabajo!

¡GRACIAS POR SU ATENCIÓN!