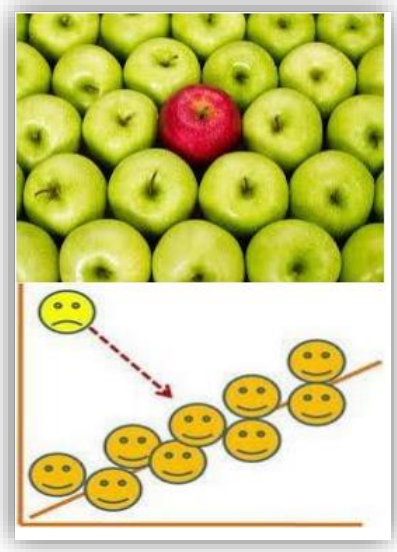
GFPI-F-135 REALIZA EL PROCESO DE LIMPIEZA DE DATOS
DETECCIÓN DE VALORES ATÍPICOS

ACTIVIDADES POR DESARROLLAR:

1. **¿Qué es un valor atípico?**
2. **Aplicación**
3. **Causas de los valores atípicos**
4. **Tipos de valores atípicos**
5. **Métodos estadísticos para detectar valores atípicos**
6. **Técnicas utilizadas**

¿Qué es un valor atípico?



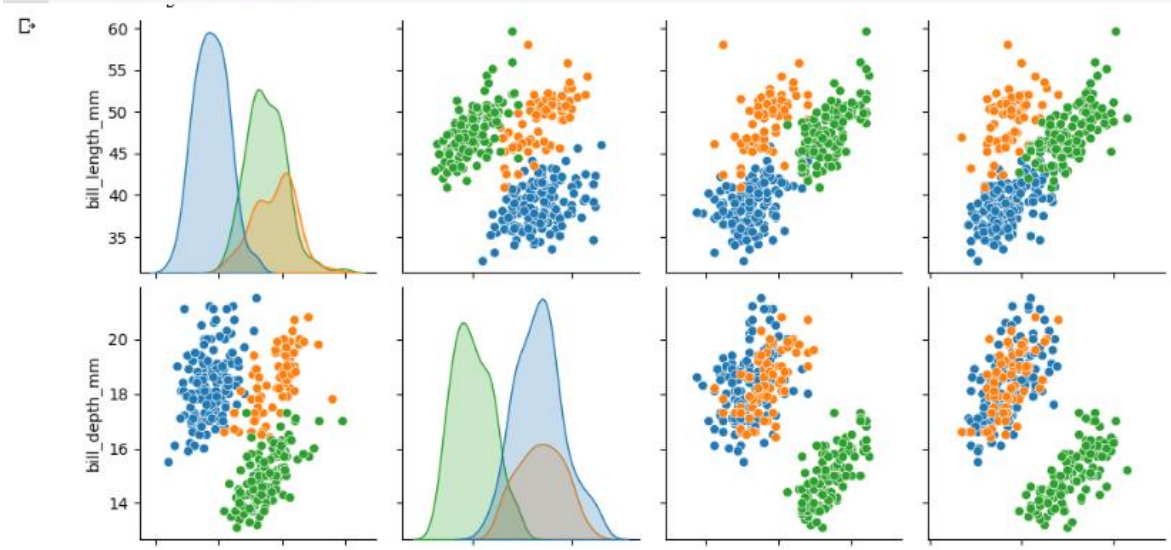
Un valor atípico (**outlier**), es un objeto de datos que se desvía significativamente de los objetos normales, valores que distan mucho del resto de conjunto de datos. como si fuera generado por un mecanismo diferente, hay que tener en cuenta:

- Los valores atípicos son diferentes de los datos de ruido
- El ruido es error o varianza aleatoria en una variable medida.
- El ruido debe ser removido antes de la detección de valores atípicos
- Los valores atípicos son interesantes: viola el mecanismo que genera los datos normales

Por ejemplo, si tenemos la serie de datos [1, 3, 5, 2, 79, 4, 8, 6], el número 79 es claramente un valor atípico. Porque su valor es extremadamente más grande que el resto de datos.

Normalmente, los valores atípicos se distinguen fácilmente en los diagramas de dispersión, ya que están aislados respecto al resto de datos.

```
import seaborn as sns
df = sns.load_dataset("penguins")
sns.pairplot(df, hue="species")
#sns.boxplot(df_boston['DIS'])
```



Aplicación

- Detección de fraudes en tarjetas de crédito.
- Detección de fraudes de telecomunicaciones
- Segmentación de clientes
- Análisis médicos

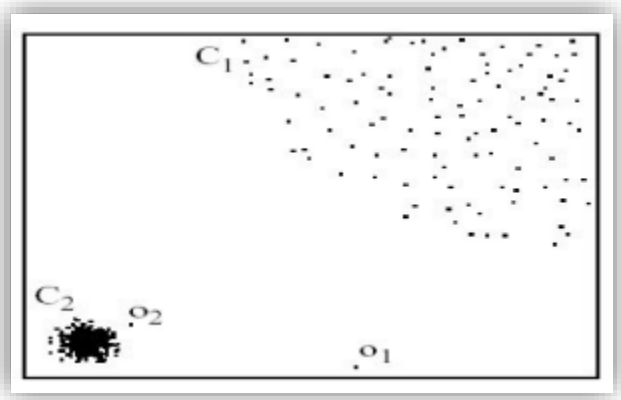
Causas de los Outliers

- Pobre calidad de los datos / contaminación
- Mediciones de baja calidad, equipo que funcione incorrectamente, error manual
- Datos correctos pero excepcionales

Tipos de Outliers

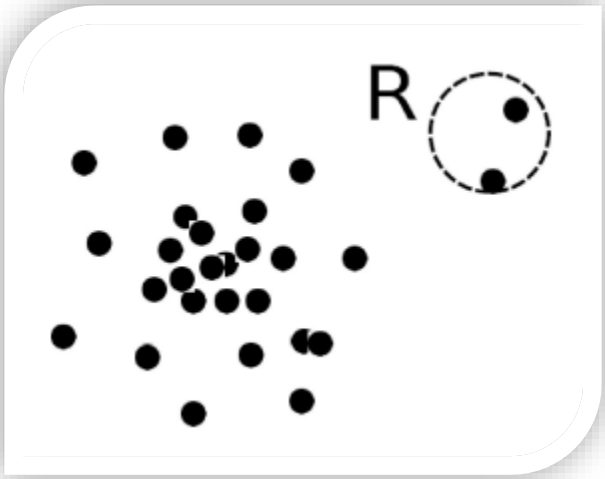
- La contaminación puede producirse por "columna", no por fila
- ¿Por qué no modificar el agrupamiento (**clustering**) u otros algoritmos para detectar los valores atípicos?

  Sólo ciertos atributos pueden tener propiedades atípicas, no es necesario descalificar la totalidad de la tupla.
- Outlier global (o anomalía de punto)
- Outlier contextual (o outlier condicional)



Métodos de estadística

También conocidos como métodos basados en modelos, suponen que los datos normales siguen algún modelo estadístico (un modelo estocástico) Datos univariados: un conjunto de datos que involucra solo un atributo o variable. A menudo suponga que los datos se generan a partir de una distribución normal, aprenda los parámetros de los datos de entrada e identifique los puntos con baja probabilidad como valores atípicos



$$\ln \mathcal{L}(\mu, \sigma^2) = \sum_{i=1}^{n} \ln f(x_i|(\mu, \sigma^2)) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Tomando derivadas con respecto a μ y σ2, obtenemos las siguientes estimaciones de verosimilitud máxima

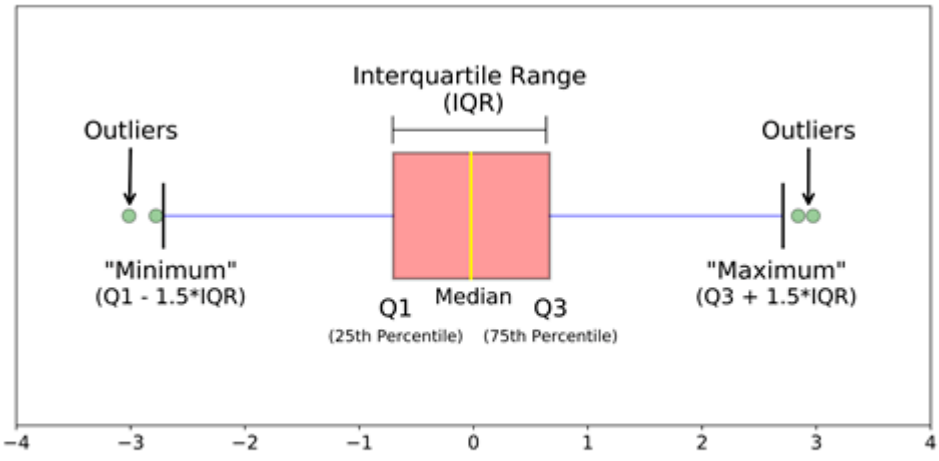
$$\hat{\mu} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Técnica para detectar outlier

Antes de comenzar a detectar valores atípicos se recomienda ver los siguientes videos: " OUTLIERS O VALORES ATÍPICOS. Teoría. Conceptos básicos" y "CÓMO calcular valores ATÍPICOS"

Los datos que vamos a trabajar se encuentran en "Dataset de trabajo", aunque primero se hará una practica en ambiente junto con el instructor llamado "Como identificar valores atípicos con Excel", que ayudará a entender el concepto de valores atípicos.

Detección de valores atípicos utilizando el rango entre cuantiles (IQR)



IQR para detectar valores atípicos

Criterios: los puntos de datos que se encuentran 1,5 veces el IQR por encima de Q3 y por debajo de Q1 son valores atípicos.

Pasos:

- Ordene el conjunto de datos en orden ascendente
- calcular el primer y tercer cuartiles (Q1, Q3)
- calcular IQR = Q3-Q1
- calcular límite inferior = (Q1−1.5 * IQR), límite superior = (Q3 + 1.5 * IQR)
- recorrer los valores del conjunto de datos y verificar aquellos que caen por debajo del límite inferior y por encima del límite superior y marcarlos como valores atípicos

Código Python:

```python
import numpy as np
sample = [15, 101, 18, 7, 13, 16, 11, 21, 5, 15, 10, 9]
outliers = []
def detect_outliers_iqr(data):
    data = sorted(data)
    q1 = np.percentile(data, 25)
    q3 = np.percentile(data, 75)
    # print(q1, q3)
    IQR = q3-q1
    lwr_bound = q1-(1.5*IQR)
    upr_bound = q3+(1.5*IQR)
    # print(lwr_bound, upr_bound)
    for i in data:
        if (i<lwr_bound or i>upr_bound):
            outliers.append(i)
    return outliers# Driver code
sample_outliers = detect_outliers_iqr(sample)
print("Outliers from IQR method: ", sample_outliers)
```

```
Outliers from IQR method:  [101]
```

Manejo de outliers

- Recortar / eliminar el valor atípico
- Revestimientos y pavimentos a base de cuantiles
- Imputación media / mediana

EVIDENCIA(S) A ENTREGAR:

1. Detectar valores atípicos utilizando un dataset propuesto por el instructor.

CONTROL DEL DOCUMENTO

	Nombre	Cargo	Dependencia	Fecha
Autor (es)	José Fernando Galindo Suarez	Instructor	CGMLTI-Teleinformática	16/02/2023

CONTROL DE CAMBIOS (diligenciar únicamente si realizan ajustes al taller)

	Nombre	Cargo	Dependencia	Fecha	Razón del Cambio
Autor (es)					