Universidad de los Andes | Educación Continua
Vicerrectoría Académica

Calidad de los datos

- Datos atípicos
- Datos faltantes
  - Supuestos
  - Tratamiento

Al final de la clase de hoy

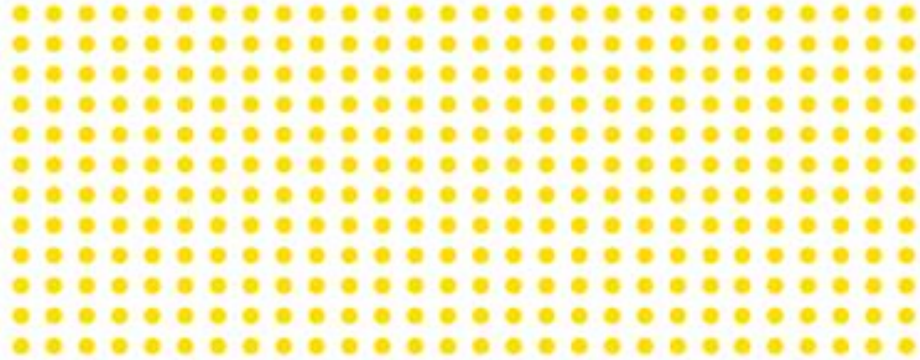
1. Estaremos en capacidad de **analizar las implicaciones** de tener valores atípicos y vacíos.

2. Sugeriremos **caminos de procedimiento** para estas situaciones con criterio.

3. Identificaremos las **limitaciones** de nuestro análisis.

Datos atípicos

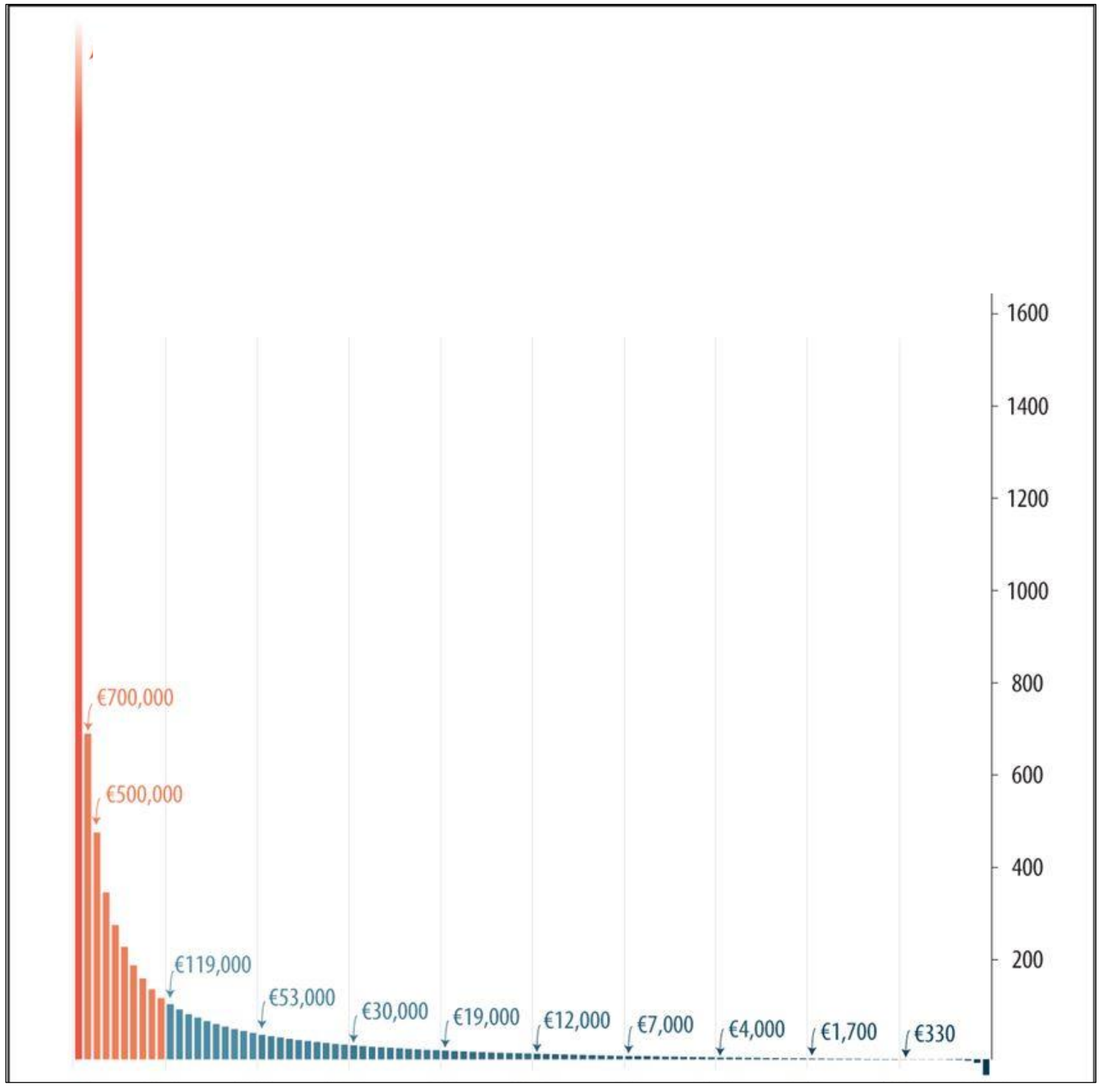
¿Qué hacer cuando hay datos con un comportamiento “diferente”



Pregunta

- Observe este gráfico sin contexto.
- Nos muestra una distribución de unos datos por percentiles.
- Hay datos negativos.
- La mayoría de datos están por debajo de 700.000.
- ¡El percentil 99 Se sale de la gráfica!

¿Usamos esos datos?

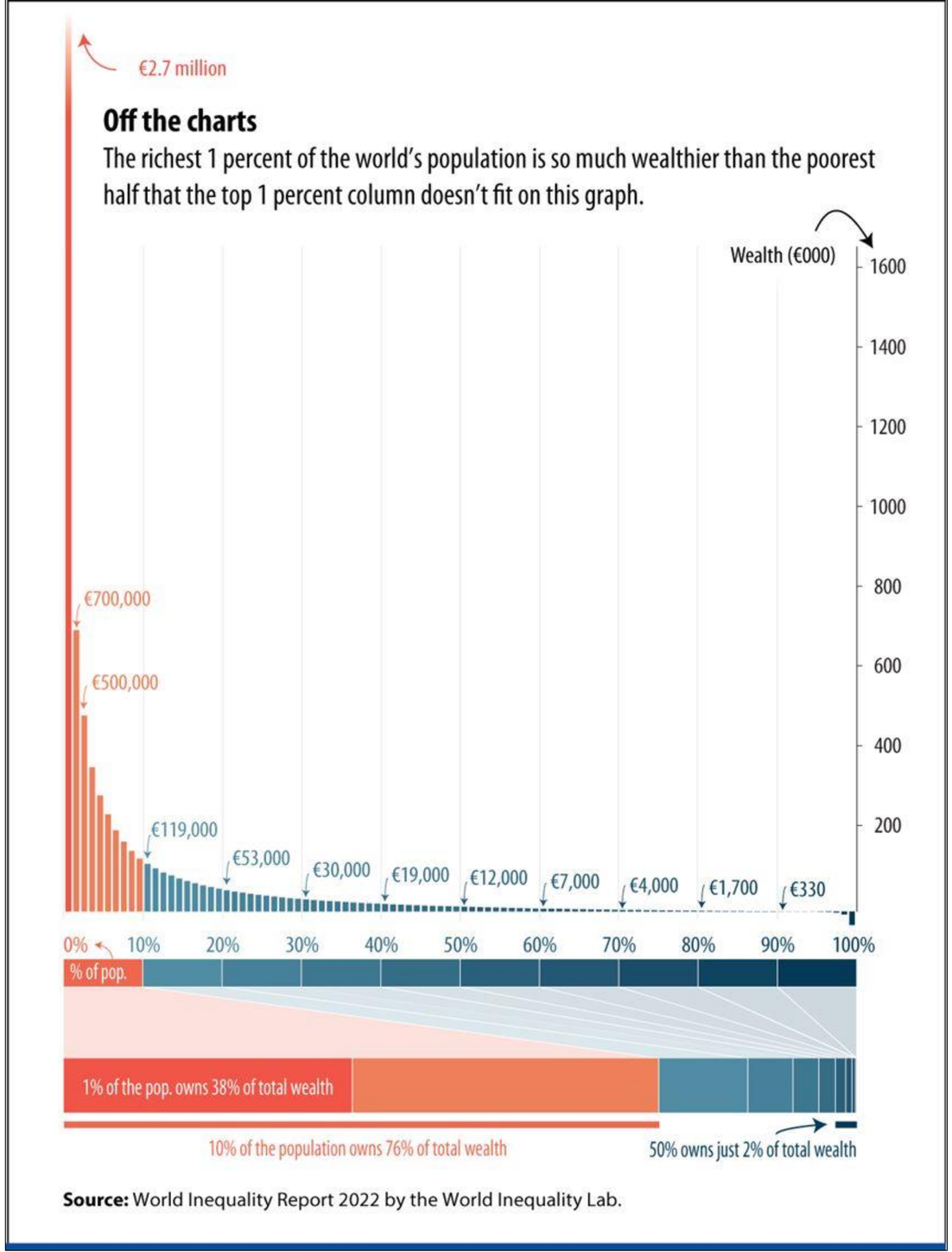


Tomen 2 minutos para escribr en el chat y hablamos

Vamos a recuperar sus ideas y volvemos al final sobre esta pregunta

Volvemos

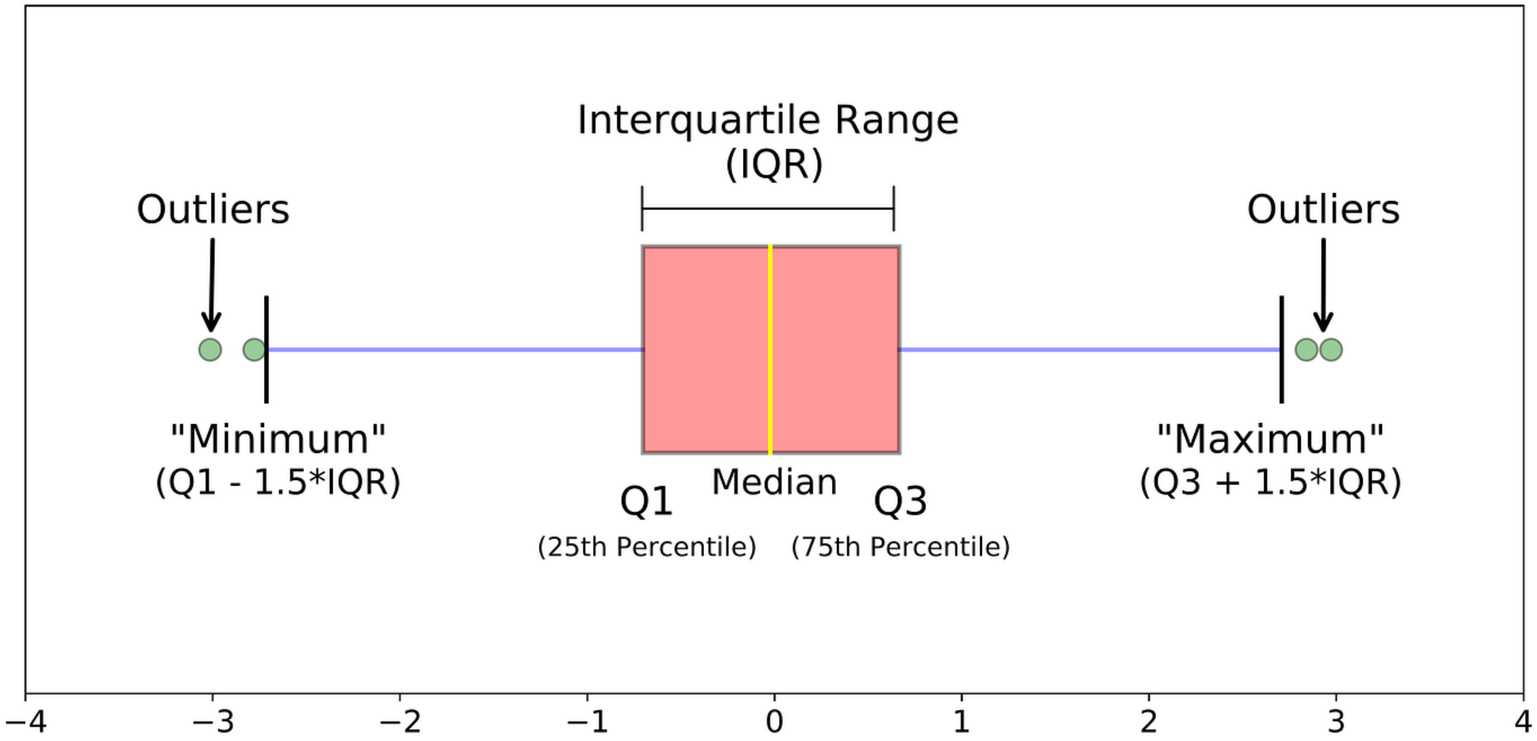
- La gráfica que observamos al principio trata sobre distribución de ingresos.
- “Los datos atípicos” de hecho, comunican el mensaje central: el 1% de las personas ganan notablemente más que las demás personas.
- Este es un patrón sistemático, no de los datos puntales. Para el propósito de este análisis debe considerarse.



Cómo se ven los *outliers*

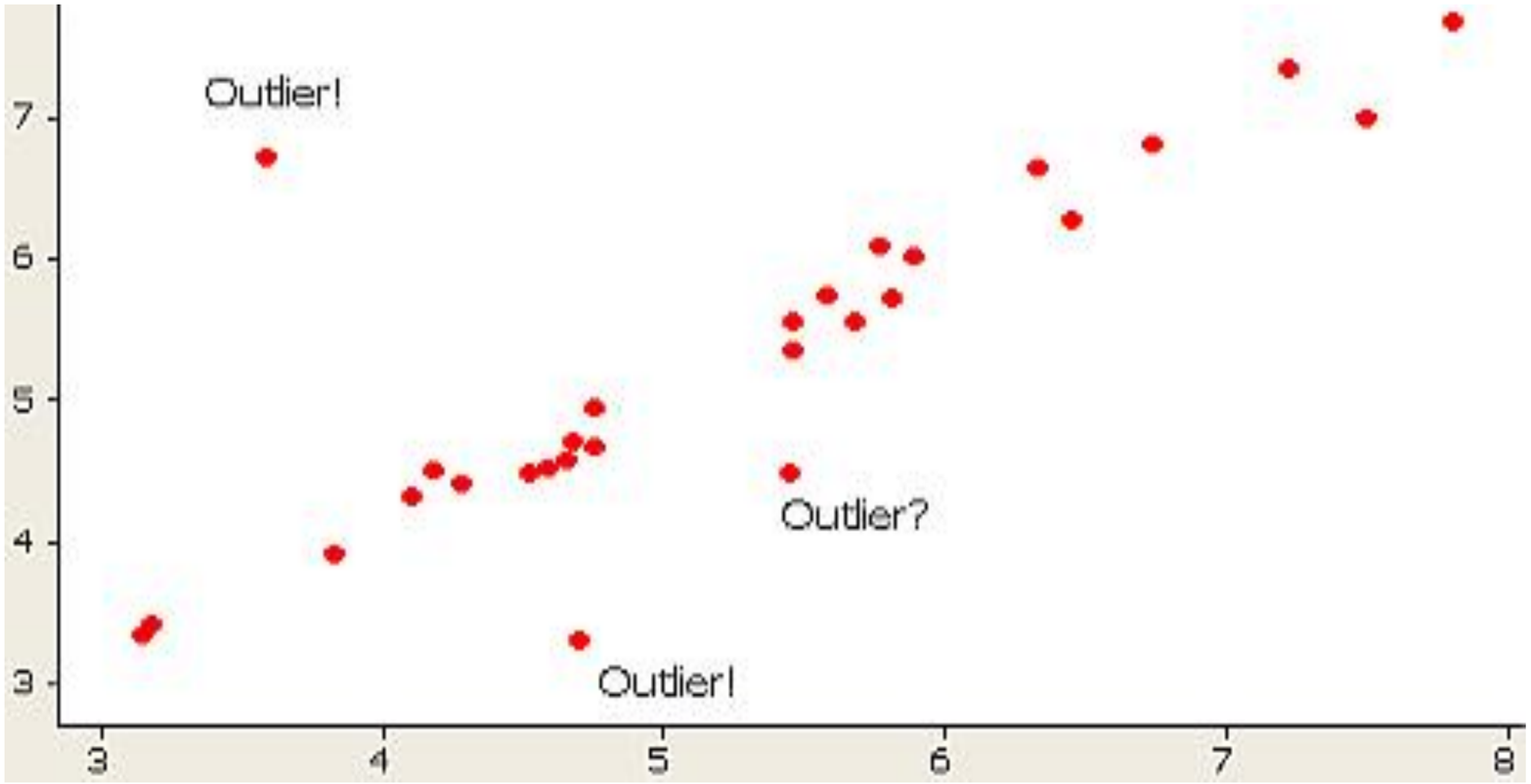
- Para reconocer datos atípicos podemos identificar patrones extraños al visualizarlos.
- Nos fijamos directamente en la dispersión de los datos para identificar visualmente si
- Hemos visto algunos gráficos en los que los valores atípicos son muy visibles.

Caja, dispersión



Un diagrama de cajas nos muestra como *outliers* datos por fuera de 1.5 veces el rango entre cuartiles 1 y 3.

Caja, dispersión



Un diagrama de dispersión nos permite también identificar visualmente posibles valores atípicos.

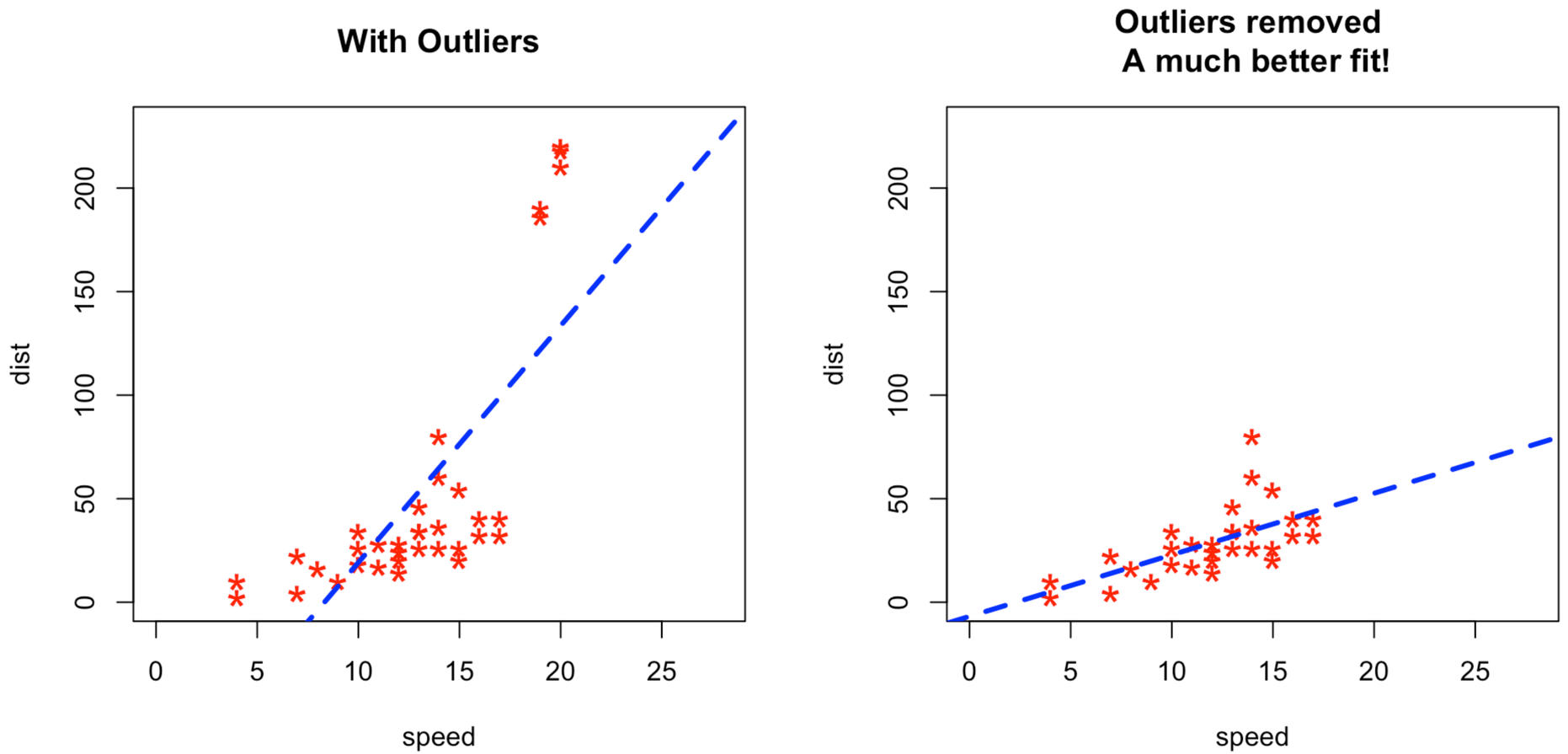
Atipico

Implicaciones

Los datos atípicos afectan fundamentalmente los estadísticos:

- Las correlaciones
- Las medidas de tendencia central (como el promedio)
- Las medidas de distribución pueden verse afectadas

Decidir cómo emplear datos atípicos puede afectar las conclusiones de un análisis.



Relación entre dos variables **con y sin** *outliers*.

Cuando veamos datos atípicos

Clave considerar: ¿Por qué observamos estos datos?

- ¿Es un tema de **limitación en la medición**?

(El sensor de medición se vuelve loco cuando llueve)

- ¿Es un tema de de una observación válida pero **sistemáticamente diferente** a las demás?

(Este punto está muy cerca de un punto de actividad volcánica)

- ¿Se trata de un **patrón sistemático** presente en otros estudios?

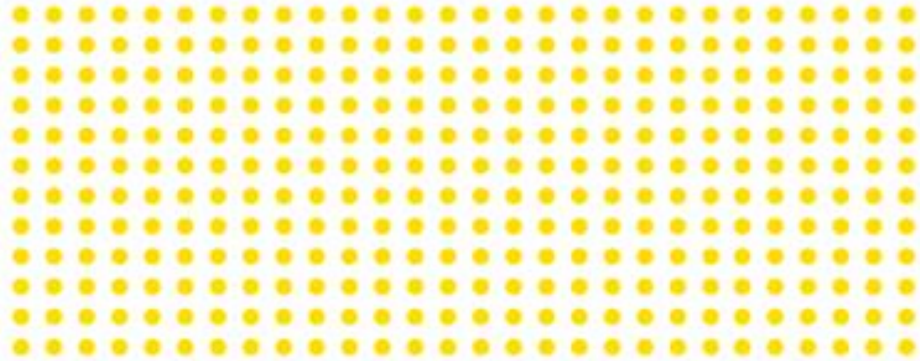
(ver diapositiva siguiente)

Cuando lo atípico es lo deseado

- En ocasiones justamente el “dato atípico” es el que queremos identificar (campo de detección de anomalías):
  - Detección de fraude financiero:
    - Transacciones atípicas de hecho comunican algo sobre un cambio en el comportamiento del usuario. Podría ser fraude.
  - Resolución de entidades
    - Queremos identificar a una entidad a través de múltiples observaciones (trabajador de alto desempeño a través de múltiples registros no identificados).
- Estas son situaciones que ameritan la aplicación de técnicas que están mas allá del alcance de esta clase.

Missing data

Lidiar con lo que no se ve…



Pregunta

Niño	Temperatura
Carlitos	
Rosita	30.4
María	
Julieta	30.2
Tomy	33.6
Julián	
John	32.5

- Esta es la lista de los niños del curso transición.
- A todos los niños que van a la primera clase se les toma la temperatura.
- ¿Cuál es la temperatura promedio del curso?

Tomen 2 minutos para escribr en el chat y hablamos

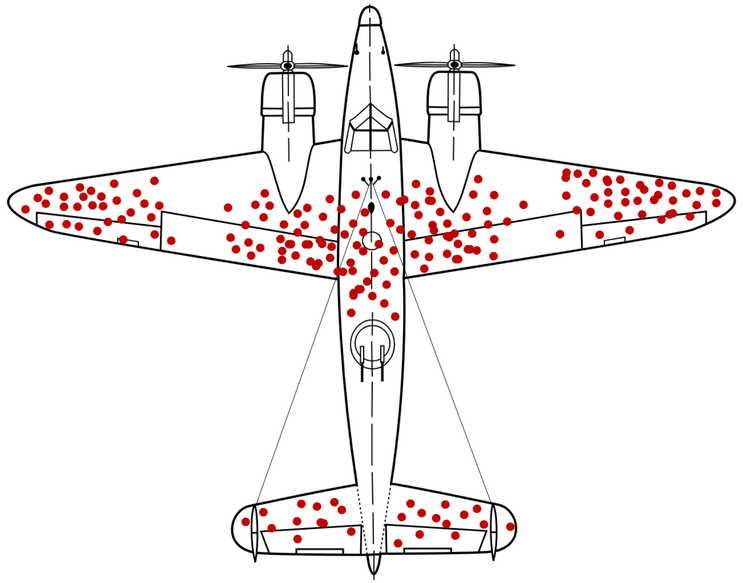
Pregunta

Niño	Temperatura
Carlitos	40.1
Rosita	30.4
María	37.9
Julieta	30.2
Tomy	33.6
Julián	39.2
John	32.5

- La temperatura promedio de los niños que vinieron es **31,6**. Ah.. muy bien…
- ¡Salvo que los datos que no observamos son precisamente de los niños que no vinieron porque se quedaron en casita con fiebre!
- Esto elevaría la temperatura promedio del curso a: **34,8** si los observáramos.

Potenciales implicaciones

- **Sesgos:** si ciegamente utilizamos sólamente la información que observamos.
- **Impresiciones en las estimaciones:** menos datos, menor tamaño de muestra para trabajar.

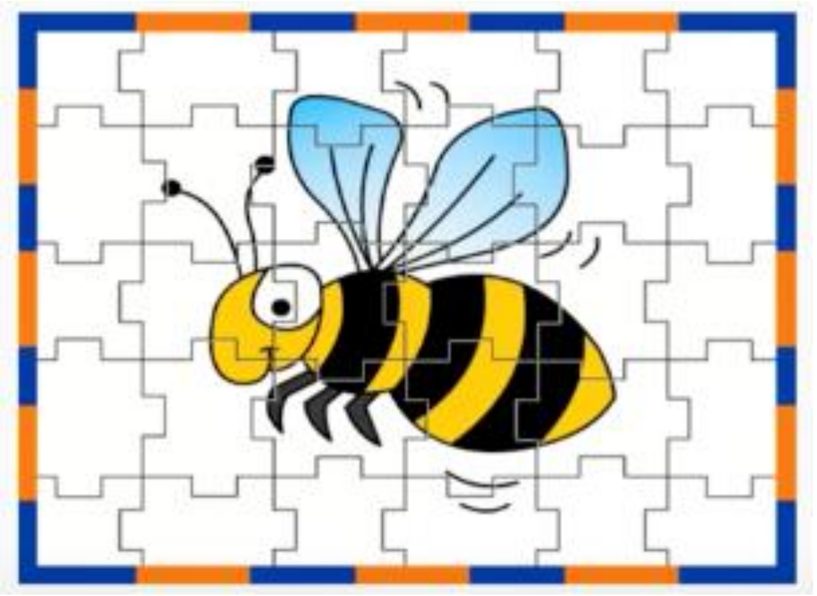


Ej. Basados en los disparos que le dieron al avión podríamos erróneamente concluir que hay que reforzar solamente la cola y las puntas del avión (pero no observamos los disparos que no lograron pegar al avión).

Tenemos los siguientes datos

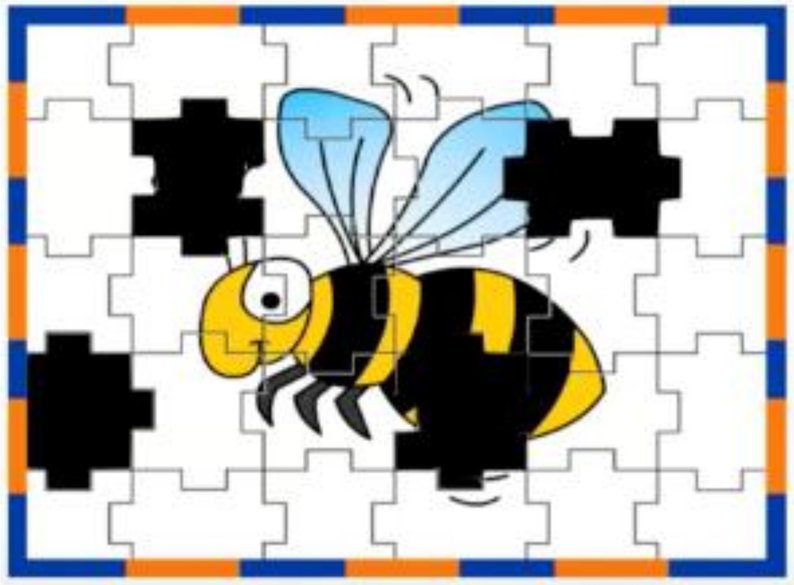
Nuestros datos se componen del color más frecuente de cada pieza del rompecabezas.

Un día un gatito se levanta y quita piezas. Analizamos la probabilidad de que una pieza falte.



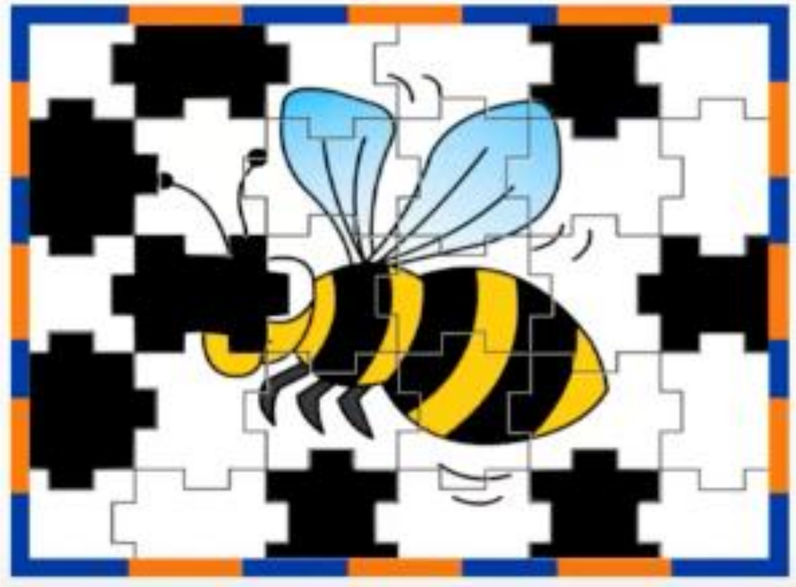
Supuestos: MCAR

- *Missing **completely** at random*
- Hay valores faltantes porque el un gato fue soltando piezas espontáneamente del rompecabezas.
- Los valores faltantes no dependen de los datos.



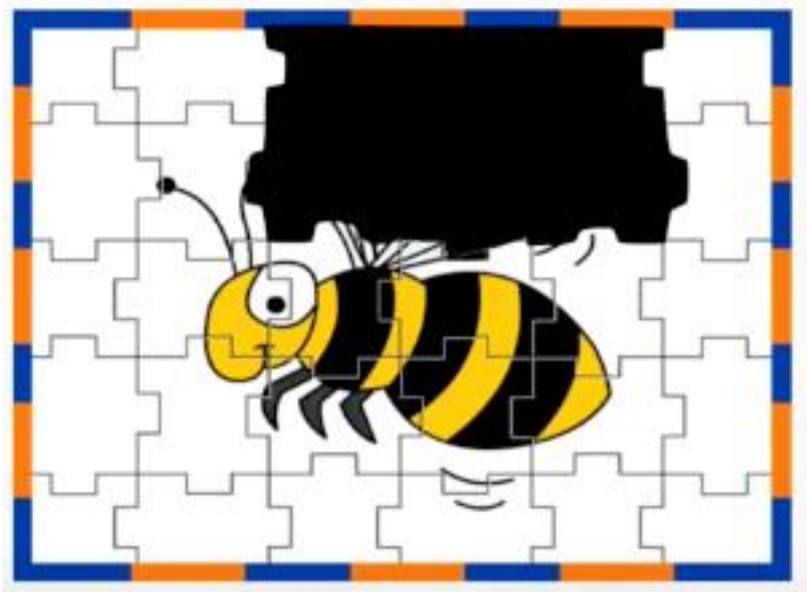
Supuestos: MAR

- *Missing at **random***
- Hay valores faltantes pero el gato está perezoso, entonces sólo quitó fichas del borde.
- Los valores faltantes dependen de si la ficha está al borde, o no.



Supuestos: MNAR

- *Missing **NOT** at random*
- El gato quitó sólo las piezas con valores azules porque no le gusta ese color
- Los valores faltantes dependen completamente de los datos.



Ejercicio

- Un reloj digital registra el número de pasos de la persona al día.
- Tenemos días en los que no hay datos.

¿Qué tipo de datos faltantes son estos? ¿Por qué?

Aleatorio, completamente aleatorio, o no aleatorio.

Tomen 2 minutos para escribr en el chat y hablamos

Depende la razón

- Si la persona se queda en cama aleatoriamente… Sería completamente aleatorio (no muy probable).
- Si los datos faltan porque el aparato se queda sin batería, probablemente los días más extensos queden sin dato (**aleatorio**).
- Si la persona se queda en cama cuando llueve, pero llueve aleatoriamente sería **completamente aleatorio**.

Tratamiento: casos completos

- Una opción es sólo usar datos que tienen las filas completas.
- Esto sólo es apropiado cuando tenemos valores faltantes **completamente aleatorios**. De lo contrario podemos introducir sesgos.
- Otra desventaja: cuando eliminamos una fila entera porque le falta dato en una columna, desechamos información valiosa en otras columnas.

Tratamiento: imputación de media

- Rellenamos los valores faltantes con el promedio de los datos para esa columna.
- **También,** sólo es apropiado cuando tenemos valores faltantes **completamente aleatorios**. De lo contrario podemos introducir sesgos.
- No estamos incorporando **incertidumbre de no observar los datos** en las observaciones. Sólamente ponemos el (único valor) promedio de la columna.

Tratamiento: imputación múltiple

- Existen formas en las que podemos hacer “adivinanzas informadas” para reemplazar los valores que faltan, según lo que conocemos de las demás características de la observación.
- En este curso no vamos a trabajar con imputación múltiple.
- Cotidianamente, a menos que estemos haciendo investigación académica, no requerimos estas técnicas de imputación.

Consideraciones

- Aunque imputar puede parecer como “hacer trampa” porque generamos “datos artificialmente”, usar ciegamente los datos completos, puede inducir a sesgos.
- Es importante considerar el tipo de aleatoriedad cuando nos enfrentamos a valores vacíos.
- No hemos visto modelos aún, pero es bueno saber que podemos predecir valores de una columna a partir de información de otra columna, incorporando incertidumbre.

Con todo esto…

1. Tenemos criterios para argumentar qué hacer con un valor no presente en los datos o con un valor atípico, **dependiendo de la situación** en la que nos encontramos.

2. Podemos sugerir caminos de acción cuando nos enfrentamos a estos datos, para el **tratamiento de estos datos**.

¡Gracias!

Aprendiendo juntos a lo largo de la vida

educacioncontinua.uniandes.edu.co

Síguenos: **EdcoUniandes**

Universidad de los Andes | Educación Continua
Vicerrectoría Académica