Módulo 2: Captura de datos

Análisis de la calidad del dato. Limpieza y mejora de los datos

Data BootCamp
Cámara Comercio Bilbao

Germán Alonso Lascurain
german@campus2b.com

TABLA DE CONTENIDOS

INTRODUCCIÓN A LA LIMPIEZA DE DATOS. ¿POR QUÉ?

- La **limpieza de datos** es un proceso **fundamental** en el que se validan los registros de las bases de datos (BBDD) con el fin de **buscar errores, corregirlos y mejorar la calidad de los mismos**.
- En este proceso se **eliminan duplicados** y datos irrelevantes. Además, se buscan datos incompletos e inexactos y se modifican mediante diversas técnicas.

Mejor pocos datos buenos, que muchos datos malos.

INTRODUCCIÓN A LA LIMPIEZA DE DATOS. ¿POR QUÉ?

- Según se ha ido incrementado las nuevas fuentes y sistemas de origen de datos, la calidad y la precisión de los mismos se ha ido degradando significativamente.

- Hay que tener en cuenta que la mayoría de las empresas tienen aplicaciones únicas para las distintas unidades de negocio. Por tanto, es probable que contengan datos incompletos,

INTRODUCCIÓN A LA LIMPIEZA DE DATOS. ¿POR QUÉ?

- Actualmente, se estima que cerca del 25% de los nuevos datos que se generan son imprecisos, incompletos, fragmentarios o meramente erróneos.

- Además, se ha estimado que los analistas pasan el 80% de su tiempo preparando los datos, mientras que dedican solo el 20% a buscar conocimiento.

https://blog.es.logicalis.com/analytics/data-quality-errores-comunes-en-la-gestion-de-la-calidad-de-datos

INTRODUCCIÓN A LA LIMPIEZA DE DATOS. ¿POR QUÉ?

- La baja calidad y suciedad de los datos puede incrementar la pérdida tanto de clientes como de la propia credibilidad del trabajo, desencadenando pérdidas en diferentes empresas (telecomunicaciones, consultoría, asesoría, salud y banca, entre otras).

- ¿Se os ocurre un caso en el que la baja calidad de los datos que pueda implicar pérdidas en una empresa?

INTRODUCCIÓN A LA LIMPIEZA DE DATOS. ¿POR QUÉ?

INTRODUCCIÓN A LA LIMPIEZA DE DATOS. ¿POR QUÉ? - PROCESO ETL

- **Extract, Transform** and **Load (ETL)** es el proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos y limpiarlos, y cargarlos en otra base de datos para analizar, o en otro sistema operacional para apoyar un proceso de negocio.

- Fuente: https://es.wikipedia.org/wiki/Extract,_transform_and_load

INTRODUCCIÓN A LA LIMPIEZA DE DATOS. ¿POR QUÉ? - PROCESO ETL

- Este proceso se puede subdividir en 5procesos:

  1)Extracción.

  2)Limpieza.

  3)Transformación. Recupera los datos limpios y de alta calidad y los estructura y resume en distintos modelos de análisis.

  4)Integración. Valida e integra los datos en BBDD para los distintos modelos de negocio.

  5)Actualización. Permite añadir los nuevos datos a las BBDD.

- Fuente: https://blogs.deusto.es/bigdata/herramientas-etl-y-su-relevancia-en-la-cadena-de-valor-del-dato/

INTRODUCCIÓN A LA LIMPIEZA DE DATOS. ¿POR QUÉ? - IMPACTO MALA CALIDAD DATO

- Una mala calidad de nuestros registros puede generar informes erróneos y defectos en el análisis, dañando así nuestra relación con el cliente ya que nos genera una dificultad de poder ofrecer un buen servicio y un trato personalizado.
- También, se crea la imposibilidad de detectar fraudes, sobre pagos,... porque no puede identificar duplicados, households, etc.

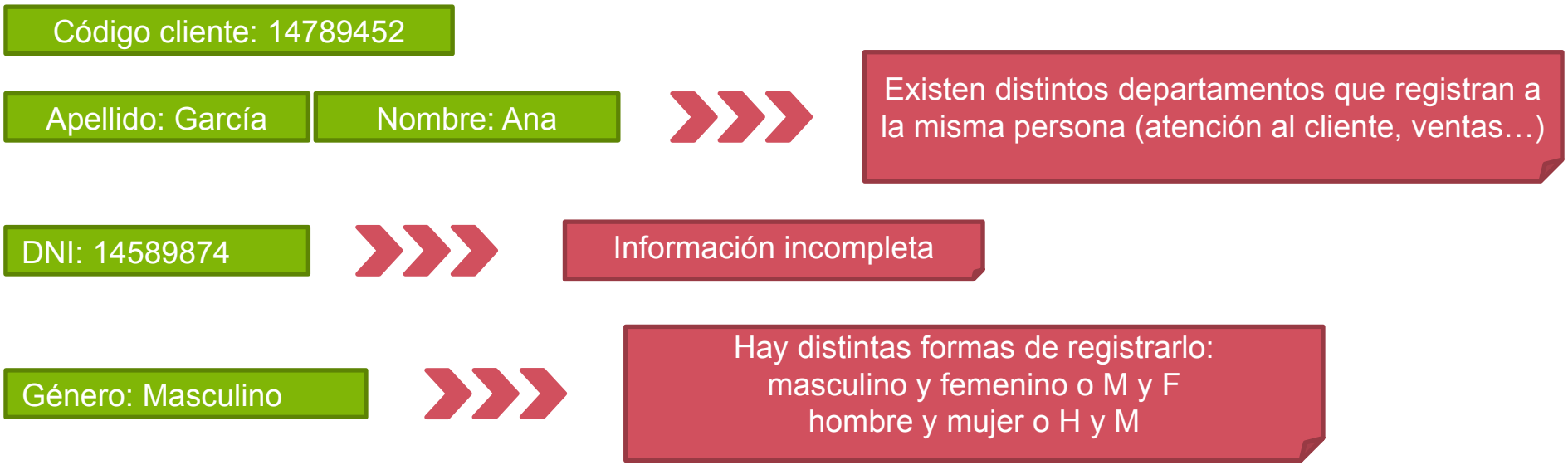
INTRODUCCIÓN A LA LIMPIEZA DE DATOS. ¿POR QUÉ? - IMPACTO MALA CALIDAD DATO

- Esta mala calidad puede llevar al incumplimiento de normativas (regulaciones y leyes).
- Además, se incrementan los costes de la gestión creando diferencias entre aplicaciones donde serán necesarios trabajos de arreglo de los registros.

TABLA DE CONTENIDOS

MOTIVOS DE IMPUREZA EN LA INFORMACIÓN

- Situaciones comunes que generan suciedad



MOTIVOS DE IMPUREZA EN LA INFORMACIÓN – CAUSAS DE LA SUCIEDAD

1. Errores de gramática. Anna o Ana.
2. Distintas unidades de medidas. Kg o g.
3. Ausencia de estándares. PC o ordenador.

MOTIVOS DE IMPUREZA EN LA INFORMACIÓN – CAUSAS DE LA SUCIEDAD

4. Diferentes formatos para un mismo campo. 12/10/2015 o 10/12/2015.
5. Valores ficticios. La altura de una persona de 456 metros.
6. Identificadores no únicos. En una sucursal un cliente con 2 identificadores.

MOTIVOS DE IMPUREZA EN LA INFORMACIÓN – CAUSAS DE LA SUCIEDAD

7. Información contradictoria. Distintos campos como el ciudad (Valencia) y la comunidad (Madrid) muestran información contradictoria.
8. Información mal relacionada. Se crean distintas BBDD para almacenar información diferente, por ejemplo la de varios clientes. Pero que luego, las tablas no se pueden unir por identificadores comunes (por estar mal definidos), y fracasamos en enriquecer la información sobre los clientes.

MOTIVOS DE IMPUREZA EN LA INFORMACIÓN – CAUSAS DE LA SUCIEDAD

9)Memoria insuficiente. Cuando se integra nueva información en las BBDD se crean errores en los procesos porque no hay espacio para el almacenado.

10)Datos anticuados. No hay una actualización correcta de la información, por ejemplo de la dirección postal, implicando entregas erróneas al cliente.

MOTIVOS DE IMPUREZA EN LA INFORMACIÓN – CAUSAS DE LA SUCIEDAD

- **Para enriquecer las fuentes de datos, ES NECESARIO poder relacionar las diferentes BBDD.**
- Si no podemos relacionar los datos de una BBDD con otra, no podremos vincular la información entre ellas.
- Unos datos sucios, pueden dar lugar a problemas para relacionar los datos entre sí.

TABLA DE CONTENIDOS

PASOS PARA UNA LIMPIEZA

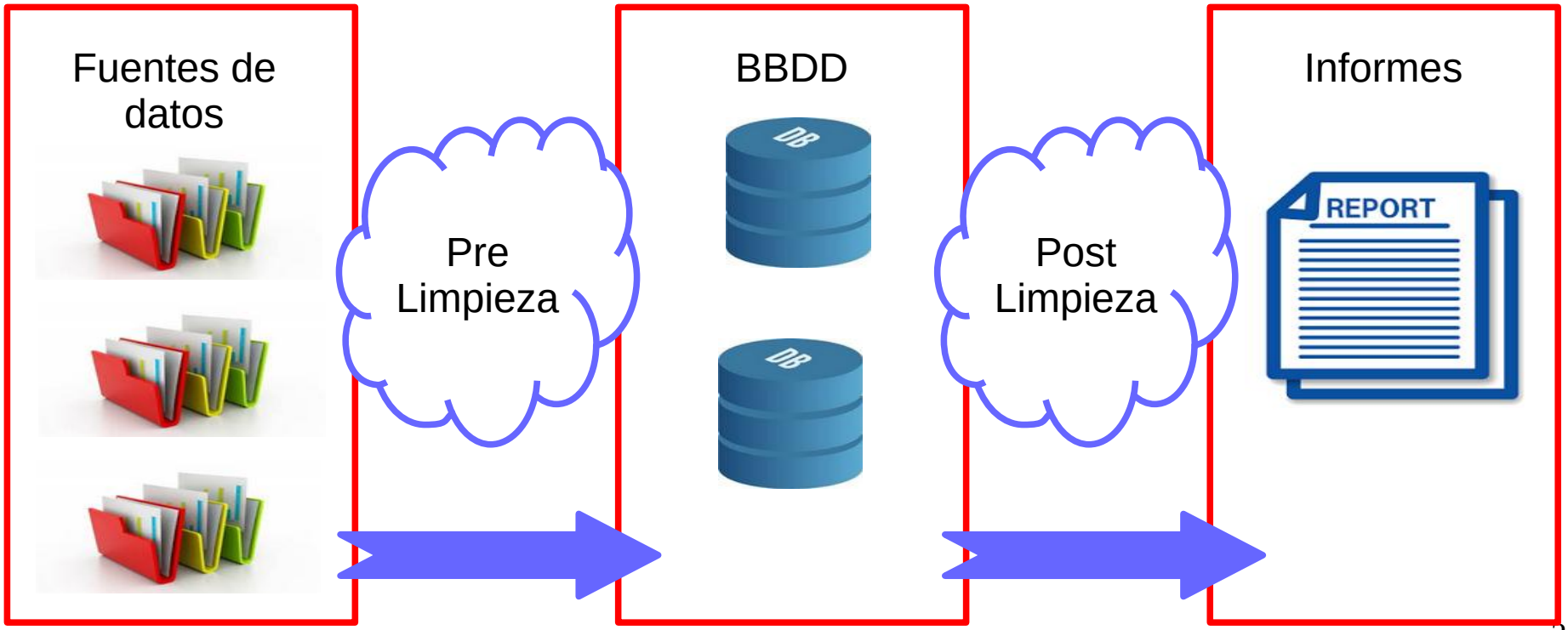
1) Pre-limpieza (solo tiene sentido para sistemas Big y es tremendamente importante para reducir los trabajos post-limpieza)

   i. Normalización o estandarización.
   ii. Determinación y separación de los datos (parsing).
   iii. Análisis de la información.

2) Post-limpieza (ya tenemos los datos en nuestras BBDD)

   i. Corrección de datos
   ii. Matching
   iii. Consolidación

PASOS PARA UNA LIMPIEZA



PASOS PARA UNA LIMPIEZA

1.i) Normalización o estandarización.

→ **formatos comunes** a **columnas comunes**.

Ej, campo fecha en 2 datasets diferentes, métricas comunes para mismas unidades (€, Kg, ...)

La mayoría de herramientas de los almacenes de datos permiten la integración de los datos en las bases en el **formato esperado**.

PASOS PARA UNA LIMPIEZA

1.ii) Determinación y separación de los datos (parsing).

→ **Descomposición** de los distintos **elementos** que componen los datos.

<u>Ejemplo:</u>
Amaia Goikoetxea Barrutia 34 Contable

Nombre: Amaia
Apellido1: Goikoetxea
Apellido2: Barrutia
Edad: 34
Profesión: Contable

PASOS PARA UNA LIMPIEZA

1.iii) Análisis de la información.

→ Determinar qué **tipo de errores** e **inconsistencias** deben ser eliminados.

Además de una **inspección manual** de las muestras de datos, es necesario la **automatización**, es decir, la incorporación de programas que actúen sobre los metadatos para detectar problemas de calidad de datos que afecten a sus propiedades.

PASOS PARA UNA LIMPIEZA

2.i) Corregir (correcting).

Consiste en **reemplazar** un elemento erróneo o vacío por uno correcto.

Hay diversos métodos para la búsqueda de errores, para el reemplazo de datos erróneos e incompletos, para la detección de valores extremos, etc. que se analizarán en el siguiente punto.

PASOS PARA UNA LIMPIEZA

2.ii) Relacionar (matching).

Se usa para descubrir campos que **emparejar** cuando se mezclan diferentes fuentes.

Tienen en cuenta un conjunto de reglas que establecen equivalencias entre los registros de diferentes bases de datos, teniendo en cuenta la combinación de varios campos a emparejar.

El Matching es FUNDAMENTAL a la hora de enriquecer los datos.

PASOS PARA UNA LIMPIEZA

2.iii) Consolidación.

Cuando se ha usado el matching para la detección de duplicados, con frecuencia se desea **fusionar** estos registros. A esto se le denomina Consolidación.

Existen dos métodos principales de consolidación:

- Registro Superviviente (nos quedamos con el datos más recientes o no).
- Mejor Registro de los duplicados.

PASOS PARA UNA LIMPIEZA

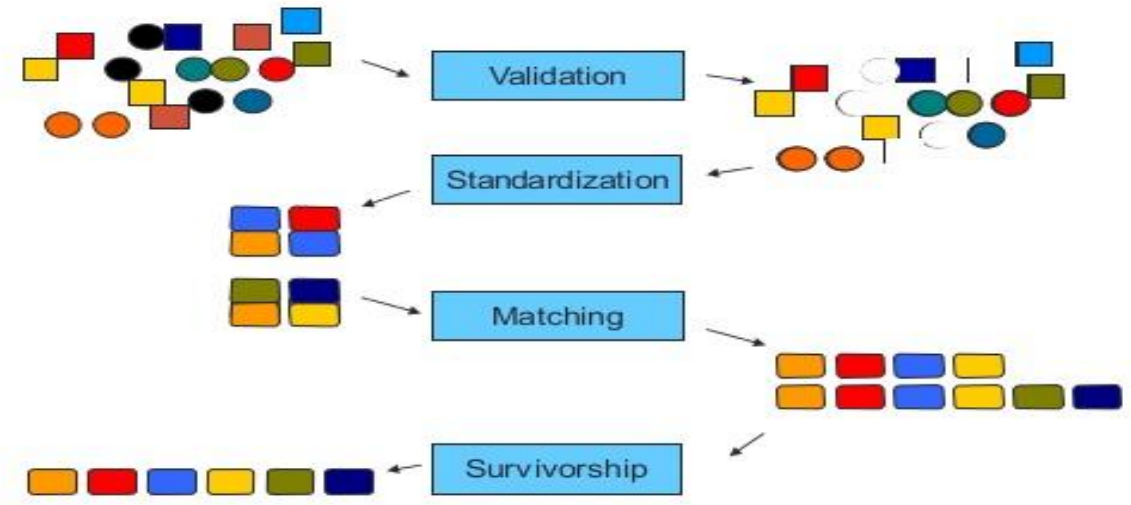


TABLA DE CONTENIDOS

MÉTODOS DE LIMPIEZA – REGLAS DE DEPURACIÓN

Las reglas de depuración nos permiten garantizar que los datos que analizamos son coherentes a la realidad.

2 tipos de reglas de depuración:

- Los edits
- Las reglas de depuración determinística

MÉTODOS DE LIMPIEZA – REGLAS DE DEPURACIÓN

Edits: Los edits **definen** las **condiciones inaceptables** o **condiciones que deben ser satisfechas** por los datos para ser rechazados y aceptados respectivamente. Se pueden considerar filtros de información.

2 tipos de reglas de depuración:

- **Los edits**, no contienen acciones correctivas (no modifican los datos del dataset). Habrá que hacerlo más adelante de manera manual o automática.
- Las **reglas de depuración determinística**, las reglas de depuración determinística, generalmente son de la forma:
  IF (condición de error) THEN (acción correctiva)

En ellas no sólo se determina la **condición** inaceptable, también se incorpora una ***solución*** para la misma.

MÉTODOS DE LIMPIEZA – REGLAS DE DEPURACIÓN

Ejemplos:

- Edit numérico

  $\mathbf{x_1/x_2 \leq 3}$ → se transforma en $\mathbf{x_1 \leq 3*x_2}$

- Edit categórico (o lógico), utiliza los operadores AND y OR

  [Edad(<15) AND (E_Civil (casado) OR Rela (CabezaFamilia))]

- Edit condicionales numéricos, tienen la forma

  IF sector = 3 THEN (PresupuesTotal / N_Empleados) < 1400

TABLA DE CONTENIDOS

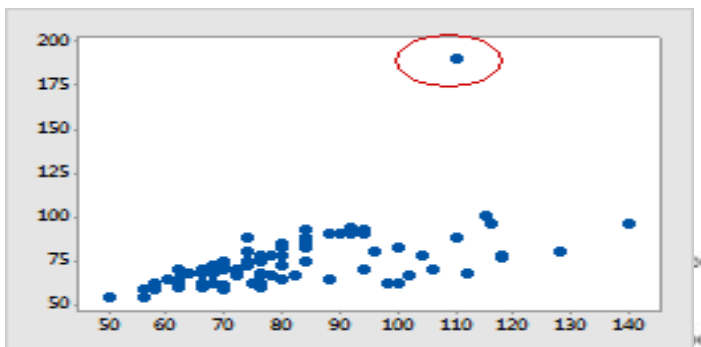
MÉTODOS DE LIMPIEZA – VALORES EXTREMOS

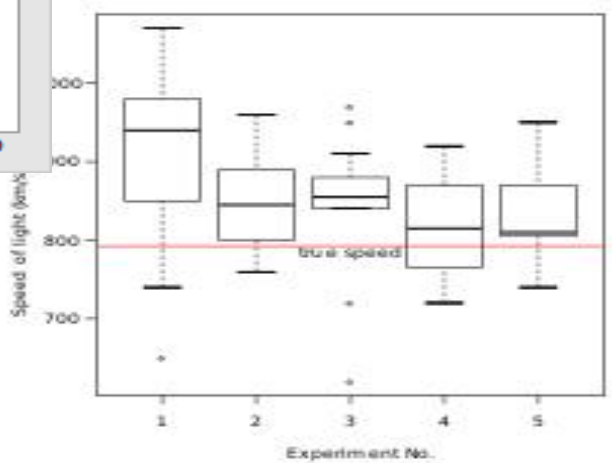
- Los valores extremos, datos atípicos o outliers, son aquellas **observaciones significativamente diferentes al resto de los datos**.
- Se refiere a aquellas observaciones que parecen ser incompatibles con el resto de los datos relativos al modelo asumido.

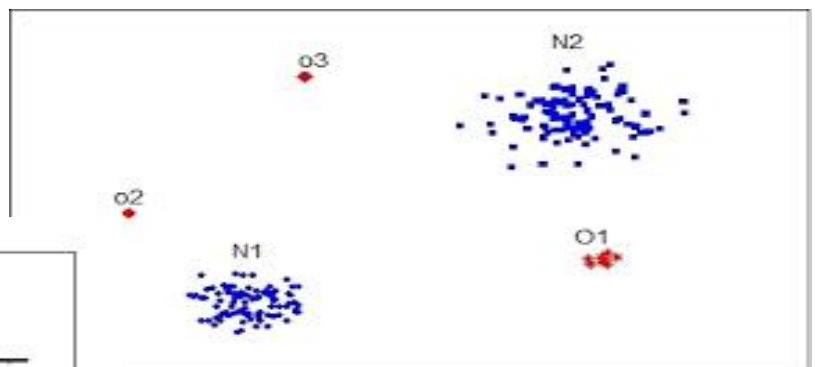
MÉTODOS DE LIMPIEZA – VALORES EXTREMOS

Podemos detectar los datos atípicos de 3 maneras:

1)Gráficamente: La representación gráfica nos permite reconocer patrones en los datos.







MÉTODOS DE LIMPIEZA – VALORES EXTREMOS

2)Detección univariante. Consideraremos valores extremos, aquellos en los que:

$$| x_i | > \bar{x} + 3*\sigma$$

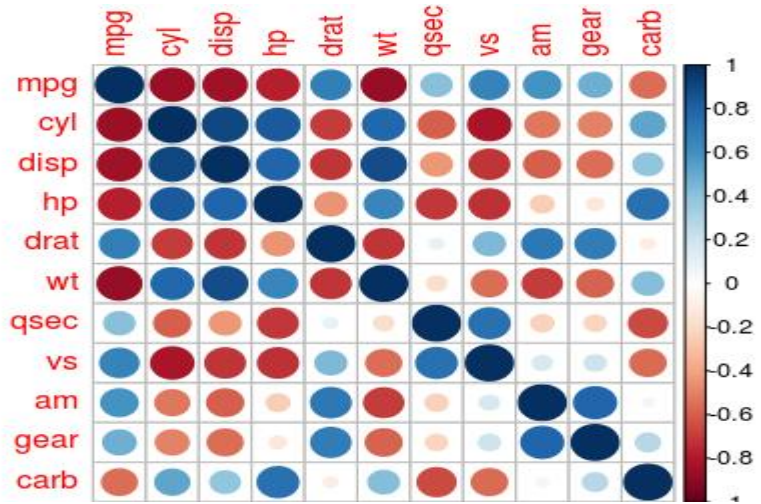
$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

MÉTODOS DE LIMPIEZA – VALORES EXTREMOS

3)Detección bivariante. Se evalúan conjuntamente pares de números.

Generalmente, para ayudar al aislamiento de los datos atípicos suele llevarse a cabo la representación gráfica de los mismos y el trazado de una elipse que determinará los valores a incluir,

el tamaño varía y nunca debe ser inferior al 50% de los datos.



MÉTODOS DE LIMPIEZA – VALORES EXTREMOS

Tratamiento de los valores extremos:

1)Podemos eliminar las filas con outliers

2)Podemos eliminar las columnas con outliers

→ Eliminar filas, implica eliminar conocimiento en la BBDD

→ Eliminar columnas, implica eliminar atributos en el análisis.

MÉTODOS DE LIMPIEZA – VALORES EXTREMOS

- Tenemos 2 maneras de proceder con los outliers:
  - Trabajar con ellos y añadir una nueva columna al dataset que nos indique que dato es un outlier.
  - Reemplazar el outlier por otro valor, con alguna de las técnicas que vamos a ver.

MÉTODOS DE LIMPIEZA – VALORES EXTREMOS

Los outliers solo se presentan en variables numéricas. Las variables categóricas no tienen outliers (ya que nosotros definimos las diferentes categorías para las variables).

TABLA DE CONTENIDOS

MÉTODOS DE LIMPIEZA – DATOS INCOMPLETOS

- Los datos faltantes aleatorios pueden **perturbar el análisis** de datos dado que disminuyen el tamaño de las muestras y en consecuencia la potencia de las pruebas de contraste de hipótesis.
- Los datos faltantes no aleatorios ocasionan, además, **disminución de la representatividad** de la muestra.

MÉTODOS DE LIMPIEZA – DATOS INCOMPLETOS

- **Localización de los datos:**

1)Se localizan los valores **faltantes o nulos**

2)Se realiza un análisis básico:

2.1)Cantidad de valores nulos, promedios, etc.

2.2)Comparar con valores esperados

Analizar información:

No hay información de ventas durante 3/1 .. 3/4 ?

No hay productos con precio > 20 ?

MÉTODOS DE LIMPIEZA – DATOS INCOMPLETOS

- **Tratamiento de los valores faltantes o nulos.**

  1)Casos completos o eliminación

  2)Selección por variables

  3)Imputación simple

  4)Imputación múltiple

MÉTODOS DE LIMPIEZA – DATOS INCOMPLETOS

- **Tratamiento de los valores faltantes o nulos.**

  1)Casos completos o eliminación:

  Consiste básicamente en quedarnos solamente con los registros y observaciones que tienen todos sus valores, desechando los que tengan algún valor nulo o faltante.

  → El **inconveniente**, es que el tamaño de la muestra puede reducirse considerablemente, afectando a la representatividad de la muestra.

MÉTODOS DE LIMPIEZA – DATOS INCOMPLETOS

- **Tratamiento de los valores faltantes o nulos.**

  2)Selección por variables: Se mantienen en la BBDD los casos, con tal que tengan **datos en las variables** que van a ser **utilizadas** para el **análisis**.

→ Este procedimiento tiene el inconveniente de generar muestras heterogéneas.

MÉTODOS DE LIMPIEZA – DATOS INCOMPLETOS

- **Tratamiento de los valores faltantes o nulos.**

  3)Imputación simple: Los métodos de imputación simple consisten en **estimar los valores ausentes** en base a los valores válidos de otras variables y/o casos de la muestra. Se puede ejecutar una sustitución por la media, imputación por regresión, moda, una constante, ...

  → Deben aplicarse con gran precaución porque pueden introducir relaciones inexistentes en los datos reales.

MÉTODOS DE LIMPIEZA – DATOS INCOMPLETOS

- **Tratamiento de los valores faltantes o nulos.**

  3.1)Sustitución por la media: Consiste en sustituir el valor ausente por la **media aritmética de los valores válidos.**

  → Dificulta la estimación de la varianza.

  → Distorsiona la verdadera distribución de la variable.

  → Distorsiona la correlación entre variables dado que añade valores constantes.

MÉTODOS DE LIMPIEZA – DATOS INCOMPLETOS

- **Tratamiento de los valores faltantes o nulos.**

  3.2)Sustitución por constante: Consiste en sustituir los valores ausentes por **constantes** cuyo valor viene determinado por **razones teóricas** o relacionadas con la **investigación previa**.

→ Presenta los mismos inconvenientes que la sustitución por la media, y solo debe ser utilizado si hay razones para suponer que es más adecuado que el método de la media.

MÉTODOS DE LIMPIEZA – DATOS INCOMPLETOS

- **Tratamiento de los valores faltantes o nulos.**

  3.3)Sustitución por regresión: Consiste en estimar los valores faltantes a través de una regresión lineal.

→ Presenta problemas de sobreestimación de asociación entre variables.

→ Aumenta el valor del estadístico R^2

MÉTODOS DE LIMPIEZA – DATOS INCOMPLETOS

- **Tratamiento de los valores faltantes o nulos.**

4)Imputación múltiple: Es la más robusta de todas las imputaciones. Básicamente crea varias copias ALEATORIAS de la BBDD original sobre las que realiza una regresión (estocástica). Posteriormente se combinan los resultados obtenidos.

→ Librería **MICE**

→ Aunque es el mejor método, en función de los datos que tengamos, el % de valores faltantes, ... otros métodos pueden ofrecer mejores resultados.

TABLA DE CONTENIDOS

MÉTODOS DE LIMPIEZA – DATOS TRUNCADOS O CENSURADOS

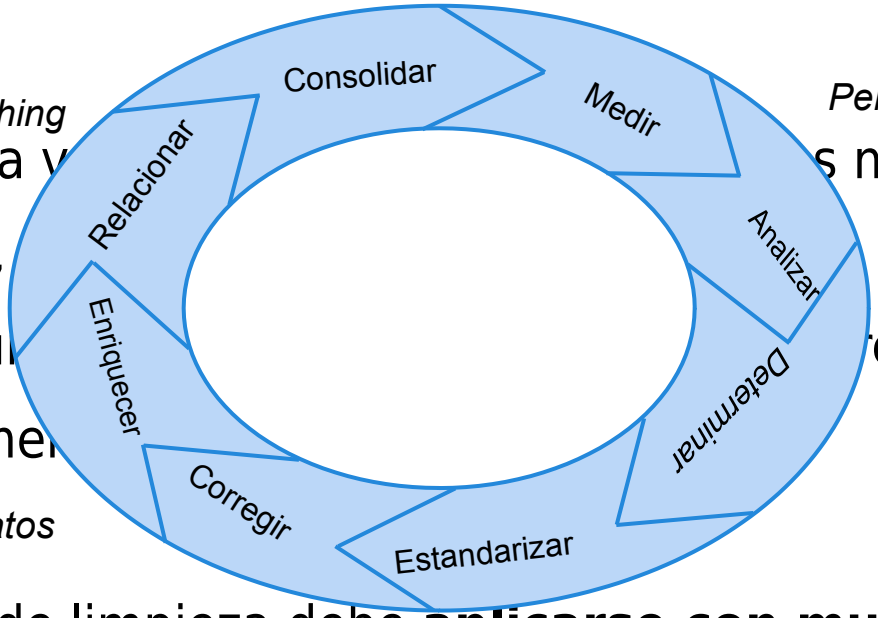
- Supongamos un estudio de ventas.

- Truncamos una variable si el valor de la venta es mayor que 500€, por ejemplo 570€, y establecemos el valor a 500€.
- Censuramos una variable si eliminamos aquellos registros cuyo valor de la venta sea menor que 0,5€.

- Este método de limpieza debe **aplicarse con mucha precaución y criterio** porque puede producir un **análisis inexacto y sesgado** (cuanto menor sesgo, mejor análisis).

MÉTODOS DE LIMPIEZA – CICLO DE CALIDAD DEL DATO

- Supongamos un estudio de ventas.



- Truncamos una v s mayor que 500€, por ejemplo 570€,
- Censuramos u egistros cuyo valor de la venta sea me
- Este método de limpieza debe **aplicarse con mucha precaución y criterio** porque puede producir un **análisis inexacto y sesgado** (cuanto menor sesgo, mejor análisis).

Germán Alonso Lascurain
german@campus2b.com
germanalonso@opendeusto.es