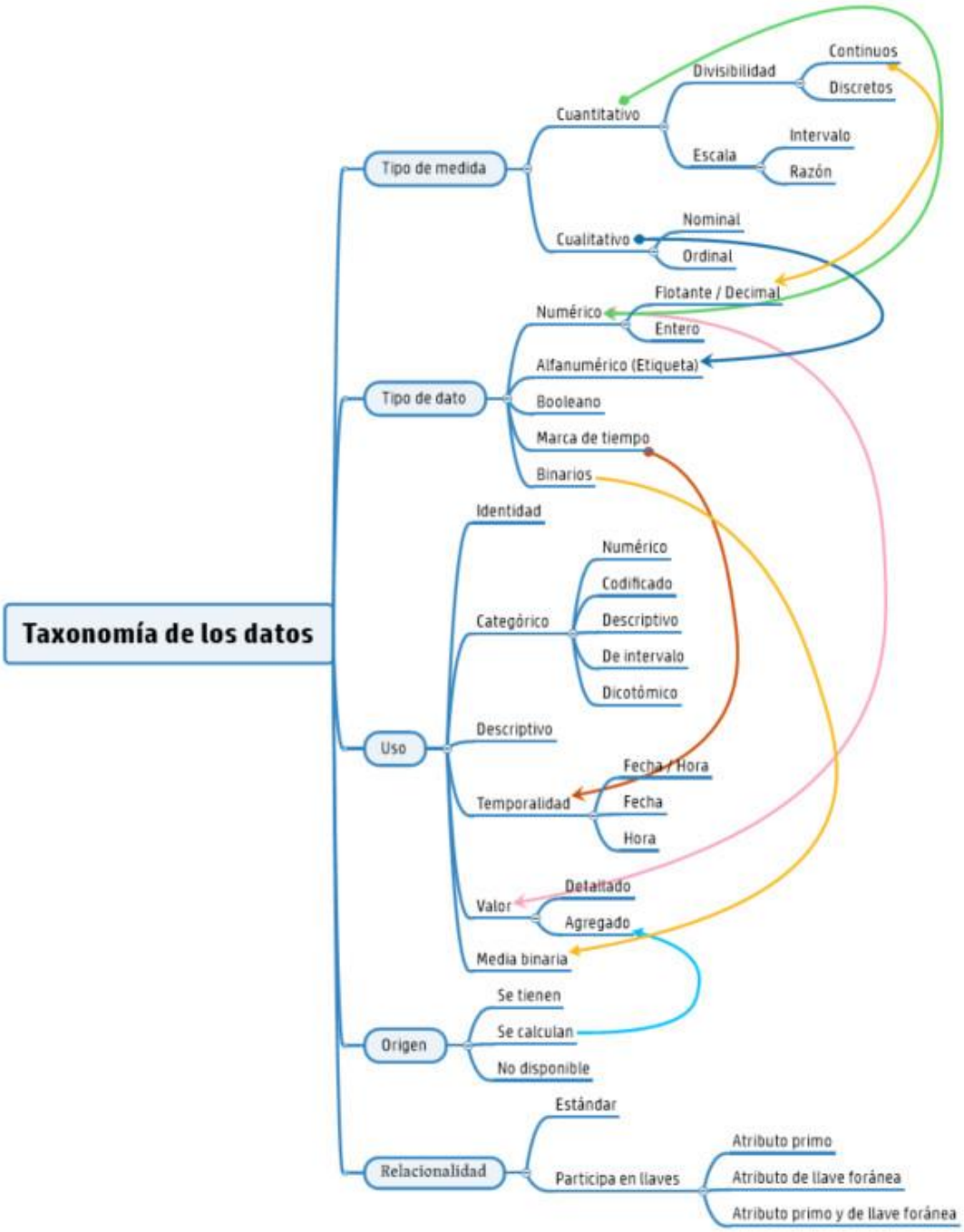
Taxonomía de los datos v4.0

Ramírez, Salazar y Araiza (2023).

Contenido

1. Mapa ilustrativo

La siguiente imagen resume el contenido de este documento.



2. Dato (definición)

Un ***dato*** es ***a)*** la representación simbólica, ya sea mediante números o letras, ***b)*** de una característica cualitativa o cuantitativa ***c)*** que permite la descripción de una cosa, suceso o estado, ***d)*** en un contexto determinado.

Los datos son muy versátiles, y es necesario conocerlos a profundidad para sacarles todo el provecho posible.

La ***taxonomía de los datos*** es un sistema de categorización que permite clasificar a los datos desde diferentes perspectivas, permitiéndonos anticipar su utilidad y sus capacidades de procesamiento; dependiendo de la naturaleza de los datos, pueden o no ser utilizados con cierta efectividad para tareas de analítica de datos.

3. Tipo de medida

En estadística, una *variable* es una característica de una muestra o población de datos, que puede adoptar diferentes valores. Las variables asumen valores dentro de una escala, entendiendo por *escala* el conjunto de valores en donde se ha de encontrar una medición; una *medida* es el dato representado en una unidad de medida de referencia, que obtenemos al medir. La medida siempre está dentro de la escala.

Considerando los **TIPOS DE MEDIDA**, aunque en algunos textos se refieran como **TIPOS DE VARIABLES ESTADÍSTICAS**, los datos pueden ser *cuantitativos* y *cualitativos*.

3.1. Datos cuantitativos

1. Los *datos cuantitativos* son características numéricas que representan cosas que se pueden medir o contar objetivamente.
    a. Ejemplos de datos cuantitativos son el `peso`, la `edad`, o un `precio`.

3.1.1. Datos continuos y discretos

2. Los *datos cuantitativos*, en cuanto a su *divisibilidad*, pueden ser *continuos* o *discretos*.

3. Se consideran *datos continuos* aquellas características que pueden dividirse para ganar más precisión, dentro de una misma unidad de medida; generalmente se tratan de mediciones que admiten decimales que pueden extenderse hasta el infinito.
    a. Ejemplos de datos continuos son la `distancia`, o el `peso`. Con la `distancia`, por ejemplo, podemos decir que algo está a 1.4 metros; pero podemos ir más a detalle, y decir que está a 1.43 metros, o a 1.4382 metros, y así.
    b. Estos datos son muy importantes para la analítica de datos, porque son los que mejor comportamiento tienen en algunos gráficos, y son requeridos por algunas técnicas estadísticas.
    c. Los datos cuantitativos continuos terminan conteniendo mediciones de características, y dada su naturaleza, debes tomar en cuenta que tienen atributos de *exactitud*, *precisión* y *error*, que no tienen otros tipos de datos.
        i. La *precisión* es el detalle con que un instrumento o procedimiento puede medir una característica.
            1. Por ejemplo, en una medición de `distancia`, una medición de 1.5 metros es menos precisa que una medición de 1.5723 metros.

        ii. La *exactitud*, por otro lado, es el grado en que la medición se acerca al valor real de lo que es medido.
            1. Generalmente, a mayor precisión es posible que se aspire a mayor exactitud. Si una persona mide 1.8734 metros, decir que mide 1.873 metros es más exacto que decir que mide 1.8 metros.

iii. Finalmente, el *error* es la diferencia entre el valor real y el valor medido.

1. Dado que es común que un dato continuo pueda seguir ganando precisión hasta el infinito, podemos decir que siempre hay error, y nunca es completamente exacto.

4. Por otro lado, se consideran *datos discretos* aquellas características que no pueden hacerse más precisos dentro de una misma unidad de medida; generalmente se tratan de conteos de naturaleza entera.
    a. Un ejemplo de dato discreto es el `número de personas`, porque no puedes decir que hay 1.5 personas para aumentar la precisión.

3.1.2. Datos de escala de intervalo y de razón

5. Los *datos cuantitativos*, en cuanto a su *escala*, pueden ser *de escala de intervalo* y *de razón*.
    a. Los *datos de escala de intervalo* son características numéricas donde la magnitud entre dos números de la escala es igual, pero la razón entre los números de la escala no es necesariamente la misma. Esto se debe a que en la escala de intervalo no se tiene una medida que implique ausencia de magnitud, conocida también como *cero absoluto*.
        i. Un ejemplo de dato de escala de intervalo es la `temperatura`: sabemos que la diferencia ente 6 grados centígrados y 7 grados centígrados, es la misma que entre 7 grados y 8 grados, pero no podemos decir que 16 grados es el doble que 8 grados.

ii. Este tipo de datos admiten operaciones aritméticas y de igualdad.

iii. Respecto a las operaciones comparativas, solo son posibles cuando se determina un cero relativo ficticio, es decir, un punto en la escala que, sin representar ausencia de magnitud, se toma como referencia para las comparaciones.

iv. Los datos de escala de intervalo, dependiendo si son continuos o discretos, pueden ser cuantitativos discretos de escala de intervalo y cuantitativos continuos de escala de intervalo.

b. Por otro lado, los *datos de escala de razón* son características numéricas donde la magnitud entre dos números de la escala es igual, y la razón entre los números es la misma gracias a que la escala tiene *cero absoluto*.

i. Un ejemplo de dato de escala de razón son las `cantidades monetarias`: sabemos que la diferencia ente 10 pesos y 20 pesos, es la misma que entre 20 pesos y 30 pesos; sabemos también que el doble de 10 pesos es 20 pesos, y sabemos que tener cero pesos equivale a no tener dinero.

ii. Este tipo de datos admiten operaciones aritméticas, comparativas y de igualdad.

iii. Los datos de escala de razón, dependiendo si son continuos o discretos, pueden ser cuantitativos discretos de escala de razón y cuantitativos continuos de escala de razón.

3.2. Datos cualitativos

6. Respecto a los *datos cualitativos* podemos decir que son características alfanuméricas que pueden observarse subjetivamente, pero no medirse.
    a. Ejemplos de datos cualitativos, son el `nombre de producto`, `color`, el `sexo de las personas`, el `nombre del mes`.

3.2.1. Datos de escala nominal y ordinal

7. Los *datos cualitativos*, en cuanto a su *escala*, pueden ser *nominales* u *ordinales*.
    a. Son *datos de escala nominal* aquellas características que representan etiquetas que identifican una categoría que no tiene un orden implícito.
        i. Ejemplo de datos de escala nominal pueden ser el `color`, el `nombre de sucursal`, el `tipo de producto`.
        ii. Este tipo de datos admiten únicamente comparaciones de igualdad.
    b. Por otro lado, son *datos de escala ordinal* aquellas características que representan etiquetas que identifican una categoría con orden implícito.
        i. Ejemplo de datos de escala ordinal son los `nombres de los meses` (Enero, Febrero, Marzo, etcétera), o el `rango militar` (Soldado, Cabo, Sargento, Capitán, etcétera). Aunque son etiquetas, debe respetarse el orden o jerarquía que implican.
        ii. Este tipo de datos admiten operaciones comparativas y de igualdad.

4. Tipo de datos

Otra tipología que podemos considerar es la relacionada con los **TIPOS DE DATOS** en general. Bajo esta clasificación, los datos pueden ser *numéricos, alfanuméricos, booleanos* y *temporales*.

4.1. Datos numéricos

1. Los *datos numéricos* son características numéricas que representan valores con los cuales se pueden hacer operaciones aritméticas.
    a. Ejemplo de este tipo de datos, son `precio`, `edad`, `cantidad`.

4.1.1. Datos flotantes / decimales y enteros

2. Los *datos numéricos* pueden ser a su vez *flotantes o decimales*, o *enteros*.
    a. Son *datos flotantes o decimales* aquellas características que tienen precisión expresada en términos decimales.
        i. Son ejemplos de datos decimales, el `peso` o la `altura`.
        ii. Es importante mencionar que cuando se manejan datos decimales en lenguajes de programación y bases de datos, suele hablarse de dos conceptos: *precisión* y *escala*, y nos referimos a la manera en que se estructura el valor con relación al punto flotante.
            1. La *precisión* sería el número total de dígitos que compone un número, por ejemplo 457.4554, tendría una precisión de 7.
            2. La *escala*, por otra parte, sería el número de dígitos que se encuentran a la derecha del punto decimal, en este caso, 4.

b. Son *datos enteros* aquellas características que expresadas en unidades indivisibles.
    i. Son ejemplos de datos enteros la `edad en años`, o el `número de hijos`. En general, hablamos de aquello que se puede contar, pero no medir.

4.2. Datos alfanuméricos

3. Los *datos alfanuméricos* son características compuestas por letras o números, y no admiten operaciones aritméticas.
    a. Ejemplo de datos alfanuméricos está el `nombre`, el `rfc`, el `código de producto`.
    b. Aquí es importante mencionar que algunos datos alfanuméricos están compuestos únicamente por dígitos, y dan la apariencia de ser números, aunque no funcionan como tal. Por ejemplo, un `número de empleado` es un identificador, y aunque es un número, debe tratarse como alfanumérico en el sentido que no es válido realizar operaciones como suma o promedio, con dicho dato.

4.3. Datos booleanos

4. Los *datos booleanos* son características que admiten el valor de falso o verdadero.
    a. Ejemplos de datos booleanos podrían ser el resultado de una comparación o una igualdad.
    b. No se les debe confundir con los *datos dicotómicos*, que son aquellos que aceptan únicamente dos valores mutuamente excluyentes, como activo / inactivo, sí / no, aceptado/rechazado, y así.

i. Los datos dicotómicos son generalmente alfanuméricos categóricos, y no representan un falso o verdadero, por lo tanto, es incorrecto considerarlos como booleanos.

4.4. Marca de tiempo

5. Los *datos de marca de tiempo* son características que representan un momento en el tiempo, conformado por una fecha y hora.
    a. Ejemplos de datos de marca de tiempo pueden ser la `fecha de nacimiento`, o la `hora de entrada al trabajo`.
    b. Aquí es importante tomar en cuenta que, en computación, los datos de marca de tiempo suelen manejarse como *números de serie*, que son un dato numérico flotante, donde la parte entera representa el día, y la parte decimal representa la hora.
        i. Para efectos prácticos, cada herramienta maneja su propio cero relativo de la escala, lo que hace posible hacer cálculos con fechas. Se trata de una fecha a partir de la cual se comienzan a contabilizar las fechas.
        ii. Por ejemplo, en el caso de Excel, el número de serie 1 representa el 01/01/ 1900.
            1. Como podrás suponer, este hecho provoca que no puedan ser manejados datos anteriores a dicha fecha.

4.5. Datos binarios

6. Finalmente, los *datos binarios* son datos no procesables de manera abstracta. Aquí entran los audios, los videos, las imágenes, y todo dato multimedia que no está representado simbólicamente por dígitos o signos interpretables por el humano a la simple lectura.

5. Tipo de uso

Atendiendo al **USO QUE TIENEN**, los datos pueden ser de *identidad*, *categóricos*, *descriptivos*, *temporales* y de *valor*.

5.1. Datos de identidad

1. Los *datos de identidad* son características que identifican a una observación como única.
    a. Ejemplos de datos de identidad son la `matrícula del alumno`, o el `código de producto`.
    b. Algunas de sus características son:
        i. No repiten valores.
        ii. Idealmente, no pueden ser datos ausentes o faltantes: cada observación debe tener su identificador, y no se puede omitir.
        iii. Si son numéricos, se asumen como etiquetas, pues no se pueden realizar operaciones aritméticas con ellos.

    c. Los datos de identidad pueden actuar de forma independiente del resto de los datos, en cuyo caso son *identificadores simples*, o actuar en conjunto con otros datos, dando lugar a *identificadores compuestos*.
        i. Un ejemplo de identificador simple podría ser el `número de empleado`. Basta ese dato para identificar a un empleado.

ii. Un ejemplo de identificador compuesto podría ser el identificador de un control de asistencia de empleados: se requiere el `número de empleado`, más la `fecha y hora` en la que asiste, para que pueda identificarse el fenómeno de forma única.

d. La principal característica de los identificadores es que tienen la propiedad de *unicidad*, es decir, no pueden repetir valores dentro de su misma serie de datos.

   i. Que tengan propiedad de unicidad no quiere decir que sean *únicos*, para ser considerados de identidad, dentro de un conjunto de datos.

   ii. Puede haber más de un dato de identidad, por ejemplo, en un registro de empleados, el `número de empleado` es de identidad, y el `número de seguridad social` también tiene todas las características de un dato de identidad; podría utilizarse para fines de identificación cualquiera de los dos datos.

   iii. La unicidad tiene que ver con los valores contenidos en cada variable; no quiere decir que solo pueda haber un solo dato identificador en un conjunto de datos.

e. Los datos de identidad son muy útiles en analítica de datos, porque facilitan el conteo y además facilitan la extracción de información dónde los agrupamientos solo involucran una observación.

5.2. Datos categóricos

2. Los *datos categóricos* son características que permiten agrupar o segmentar los datos atendiendo a categorías predefinidas.
    a. Ejemplos de datos categóricos puede ser el `sexo de las personas`, la `sucursal` donde labora un empleado, el `tipo de producto`, y así.
    b. Sus características son:
        i. Idealmente son etiquetas.
        ii. Pueden o no estar presentes en las observaciones (admiten datos faltantes o vacíos).
        iii. Es muy común que repitan valor en varias observaciones.
        iv. Deben tener un catálogo de categorías preestablecido.
    c. Los datos categóricos pueden ser de diferentes tipos:
        i. Son *categóricos numéricos*, cuando la categoría está representada por un número que debe ser interpretado para conocer el detalle de la categoría. Se recomienda que sean tratados como etiquetas, ya que no se pueden hacer operaciones con dichos números.
        ii. Son *categóricos codificados*, cuando la categoría está representada por una etiqueta codificada que debe ser interpretada para conocer el detalle de la categoría.
        iii. Los *categóricos descriptivos*, son etiquetas que describen de manera suficiente aquello que representan, sin necesidad de interpretación.
        iv. Los *categóricos de intervalo*, son cuando una escala es dividida en intervalos mutuamente excluyentes, en donde cada medida entra en una categoría determinada.
        v. Los *categóricos dicotómicos*, son cuando la clasificación solo admite dos segmentos, mutuamente excluyentes.

En la siguiente tabla se muestran ejemplos de cada uno de los tipos de categórico, suponiendo una lista de países de norte y centro américa, y de ellos, cuáles forman parte del T-MEC.

categórico_numérico	categórico_codificado	categórico_descriptivo	categórico_dicotómico
1	BZ	BELICE	NO
2	CA	CANADÁ	SÍ
3	CR	COSTA RICA	NO
4	SV	EL SALVADOR	NO
5	US	ESTADOS UNIDOS	SÍ
6	GT	GUATEMALA	NO
7	HN	HONDURAS	NO
8	MX	MÉXICO	SÍ
9	NI	NICARAGUA	NO
10	PA	PANAMÁ	NO

Los categóricos de intervalo, por su parte, se representan en forma de pares de valores. Por ejemplo, si en una toma de datos tenemos un campo llamado `edad`; podemos establecer un categórico llamado `rango_edad`, cuyas categorías sean:

`[0,18)`	De cero a dieciocho
`[18,30)`	De dieciocho a treinta
`[30,45)`	De treinta a cuarenta y cinco
`[45,65)`	De cuarenta y cinco a sesenta y cinco
`[65,100)`	De sesenta y cinco a cien

Un *rango de intervalo* es un término que se utiliza en estadística para referirse a un conjunto de valores consecutivos que se agrupan en intervalos de igual ancho, dentro de una escala.

Para definir un rango de intervalo, primero se establece el *ancho de intervalo*, que son las unidades de medida que representan el rango, y que debe ser constante en todo el conjunto de datos. Luego, se dividen los datos en intervalos de igual tamaño, donde cada intervalo abarca una cierta cantidad de valores.

Por ejemplo, si estamos trabajando con un conjunto de datos de edad de una población, podemos definir intervalos de 10 años (por ejemplo, de 0-9, 10-19, 20-29, y así).

Cuando el ancho del intervalo no es constante (por ejemplo, si la edad va de 0 a 9, y de 10 a 25, y luego de 25 a 35), ya no se les llama rango de intervalos, sino *clases*.

Tanto los rangos de intervalo como las clases, generalmente se indican con la letra **k**.

El rango de intervalo se utiliza comúnmente para crear histogramas, que son gráficos que muestran la distribución de los datos en diferentes intervalos. Los histogramas pueden ser útiles para visualizar patrones en los datos, como si están sesgados hacia un lado, si tienen una distribución normal, etc.

La notación matemática para representar rangos de intervalo utiliza los siguientes símbolos:

- **`Square Brackets []`**: Se utilizan para indicar que se trata de un intervalo cerrado, es decir, donde los límites del rango se incluyen en el rango.
- **`Rounded Brackets ()`**: Se utilizan para indicar que se trata de un intervalo abierto, es decir, donde los límites del rango no se incluyen en el rango.
- **`Coma ,`**: Se utilizan para separar el límite inferior y superior del rango de intervalo.

Puede darse el caso en que sólo uno de los límites se incluya, pero el otro no. De esa forma:

- **`(a,b]`** Intervalo semi-abierto, o intervalo semi-cerrado por la izquierda, donde el límite inferior no se incluye, pero el límite superior sí.
- **`[a,b)`** Intervalo semi-abierto, o intervalo semi-cerrado por la derecha, donde el límite inferior se incluye, pero el límite superior no.

Algunos ejemplos de notación de rangos de intervalo son los siguientes.

Rango de intervalo	Alcance
`[0,10]`	Valores mayores o iguales a cero, y menores o iguales a diez
`(0,10)`	Valores mayores a cero, y menores a diez
`[0,10)`	Valores mayores o iguales a cero, y menores a diez
`(0,10]`	Valores mayores a cero, y menores o iguales a diez

Al número de rangos de intervalos, se le conoce como *número de clases*, y suele simbolizarse con la letra **k**, así que, si una escala se divide en 5 intervalos, **k=5**

Generalmente, se debe decidir si todos los límites inferiores se incluyen, o si todos los límites superiores se incluyen, pero no se recomienda mezclar.

En aquellos casos en los que se decide no incluir el límite inferior y sí el superior, suele hacerse una excepción en el primer intervalo, para evitar que los datos iguales al valor mínimo del intervalo queden fuera.

5.3. Datos descriptivos

3. Los *datos descriptivos* son características que permiten describir de alguna manera a la observación.
    a. Ejemplos de datos descriptivos son el `nombre de una persona`, o la `descripción de un producto`.
    b. Las descripciones pueden ser únicas, pero difícilmente pueden ser utilizados como identificadores, pues no garantizan la propiedad de unicidad.
    c. Los descriptivos pueden confundirse algunas veces con los categóricos, y varían porque estos pretenden explicar o describir, y no enumerar.

d. Se diferencian de los categóricos descriptivos en que no cuentan con un catálogo preestablecido.

5.4. Datos de temporalidad

4. Los *datos de temporalidad* son características que permiten establecer un momento de referencia para la observación; este tipo de datos pueden contener la fecha, la hora, o la fecha y la hora.
    a. Ejemplo de este tipo de dato puede ser la `fecha de nacimiento`, o la `fecha y hora de salida de almacén`.
    b. Estos datos son fundamentales para las líneas de tiempo y para los análisis transversales y longitudinales.

5.5. Datos de valor

5. Están los *datos de valor*, que son características cuantitativas que permiten realizar operaciones aritméticas con ellos.
    a. Ejemplos de datos de valor podrían ser los `precios de producto`, o las `cantidades vendidas`.
    b. Estos datos son el corazón de la analítica de datos, pues en combinación con los datos categóricos y temporales, permiten enfocar segmentaciones y revelar patrones.
    c. Los datos de valor pueden ser *detallados* o *agregados*.
        i. Son *datos de valor detallados*, cuando el dato refiere a una medida inherente a una característica particular de la observación a la que pertenece.

1. Por ejemplo, en un registro de `empleados`, el dato `edad` es de valor detallado, porque cada empleado tiene su edad.

ii. Son *datos de valor agregados*, aquellos que refieren a un cálculo, y no son inherentes a una característica particular de la observación a la que pertenecen.

1. Por ejemplo, en un registro de `PRODUCTOS` pude haber un dato llamado `precio_máximo` y otro dato llamado `precio_mínimo`; imagina que dichos campos corresponden al precio máximo en que se haya vendido el producto, y el precio mínimo en que se haya vendido el producto.
2. Esto quiere decir que no es un dato inherente a una característica del producto, sino que el dato es el resultado de procesar el `precio_venta` de ese producto, en el registro de `VENTAS` del producto.

iii. Los datos de valor agregados siempre deben implicar el o los cálculos agregados que le dan origen al valor (sumatoria, máximo, mínimo, promedio, desviación estándar, etcétera).

1. En nuestro ejemplo, `precio_máximo` y `precio_mínimo`, implican los cálculos agregados de máximo y mínimo, respectivamente.

iv. Los datos de valor agregados siempre deben incluir la formulación del cálculo, y la referencia a los datos involucrados.

    1. En nuestro ejemplo, `precio_máximo` y `precio_mínimo`, dependen del contenido del dato `precio_venta`, de `VENTA`, para los registros asociados al producto que estemos analizando.

```
PRODUCTO.precio_máximo = MAX(VENTA.precio_venta)
```

5.6. Media binaria

6. Finalmente están los *datos de media binaria*, que son datos no procesables de manera abstracta. Aquí entran los audios, los videos, las imágenes, y todo dato multimedia que no está representado simbólicamente por dígitos o signos interpretables por el humano a la simple lectura. Su uso generalmente requiere interpretación y decodificación.

6. Origen que tienen los datos

Atendiendo al **ORIGEN QUE TIENEN**, los datos *se tienen*, *se calculan*, o están *no disponibles*.

6.1. Se tienen

Los *datos se tienen* cuando existen y están disponibles para su uso.

Es posible que los datos existan, pero si no podemos usarlos por la razón que sea, para efectos prácticos, es como si no los tuviéramos.

6.2. Se calculan

Los *datos se calculan* cuando no existen y no los tenemos aún, pero es posible calcularlos o inferirlos a partir de los datos y el conocimiento que sí tenemos.

Al proceso de calcular un dato, se le llama *función*, y sus resultados deben ser consistentes: los mismos datos siempre deben arrojar los mismos resultados.

6.3. No disponibles

Son *datos no disponibles* cuando no existen, no los podemos calcular, pero se pueden obtener como respuesta a una petición, ya sea a una persona o proceso.

Es posible que los datos sean proporcionados por personas, o bien por máquinas o algoritmos automatizados.

7. Relacionalidad

Cuando los datos se tienen en un modelo relacional, pueden ser clasificados atendiendo a las relaciones en las que participan dentro del modelo. Por exceder el alcance de este documento, no explicaremos los conceptos de base de datos relacionales.

Atendiendo a la **RELACIONALIDAD**, los datos *participan en llaves*, o son *estándar*.

7.1. Datos que participan en llaves

Los datos *participan en llaves* cuando son atributos primos, o cuando forman parte de una llave foránea.

Son *atributos primos* aquellos que forman parte de una llave primaria, o una llave candidata. Son *atributos de llave foránea*, aquellos que permiten la relación con la llave primaria de otra tabla.

En el caso de las llaves primarias compuestas, es posible que un atributo primo, en lo particular, también sea llave foránea, por lo cual sería un atributo *primo y de llave foránea*.

7.2. Datos estándar

Los *datos estándar* son aquellos que no son atributos primos, ni forman parte de una llave foránea.

Considerando este modelo, tendríamos estas clasificaciones:

Campo	Clasificación en cuanto a relacionalidad	Refiere a
ALMACEN.id_almacén	Atributo primo	
ALMACEN.nom_almacén	Atributo estándar	
INVENTARIO.id_almacén	Atributo primo y de llave foránea	ALMACEN.id_almacen
INVENTARIO.id_producto	Atributo primo y de llave foránea	PRODUCTO.id_producto
INVENTARIO.cantidad	Atributo estándar	
INVENTARIO.ubicación	Atributo estándar	
PRODUCTO.id_producto	Atributo primo	
PRODUCTO.nom_producto	Atributo estándar	
PRODUCTO.id_marca	Atributo de llave foránea	MARCAS.id_marca
MARCAS.id_marca	Atributo primo	
MARCAS.nom_marca	Atributo estándar	

8. Codificación Taxonómica de Datos (DTXC)

La *codificación taxonómica de datos* o *DTXC* (*Data Taxonomic Code*), es un sistema de codificación que le otorga un código a los datos, que representa su taxonomía dependiendo de:

a) el tipo de medida que tienen (*Measurement type*),

b) el tipo de dato que representan (*Data type*),

c) el uso que se les da a los datos (*Use*),

d) el origen de su disponibilidad (*Source*), y

e) su relacionalidad, en el caso de los datos que forman parte de un modelo de datos (*Relationality*).

El código DTXC se compone de la siguiente manera:

Measurement type/*Data type*/*Use*/*Source*/*Relationality*

En algunos casos, el separador de código puede variar, siendo posible utilizar el símbolo pipe line (|).

A continuación, se muestran los diferentes códigos DTXC.

CODIFICACIÓN TAXONÓMICA DE DATOS (DTXC/DATA TAXONOMIC CODE)

Clasificación	Categoría	Categoría en inglés	Código DTXC
Tipos de medida (*Measurement Type*)	Cuantitativo	*Quantitative*	**QT**
	Cuantitativo Discreto	*Quantitative Discrete*	**QT-DIS**
	Cuantitativo Continuo	*Quatitative Continuous*	**QT-CON**
	Cuantitativo Discreto de Escala de Intervalo	*Quantitative Discrete, Interval Scale*	**QT-DIS-IS**
	Cuantitativo Discreto de Escala de Razón	*Quantitative Discrete, Ratio Scale*	**QT-DIS-RS**
	Cuantitativo Continuo de Escala de Intervalo	*Quantitative Continuous, Interval Scale*	**QT-CON-IS**
	Cuantitativo Continuo de Escala de Razón	*Quantitative Continuous, Ratio Scale*	**QT-CON-RS**
	Cualitativo	*Qualitative*	**QL**
	Cualitativo Nominal	*Qualitative Nominal*	**QL-NM**
	Cualitativo Ordinal	*Qualitative Ordinal*	**QL-OR**
	Tipo de medida no definido	*Not Defined Measurement Type*	**NDMT**
Tipo de dato (*Data Type*)	Numérico	*Numeric*	**NUM**
	Numérico Flotante	*Numeric Float*	**NUM-FL**
	Numérico Entero	*Numeric Integer*	**NUM-INT**
	Alfanumérico	*Alphanumeric*	**STR**
	Booleano	*Boolean*	**BL**
	Marca de tiempo	*Time Stamp*	**TS**
	Binario	*Binary*	**BIN**
	Binario Muy Largo	*Binary - Binary Large Object*	**BIN-BLOB**
	Tipo de dato no definido	*Not Defined Data Type*	**NDDT**
Uso (*Use*)	Identidad	*Identity*	**ID**
	Categórico	*Categorical*	**CAT**
	Categórico Numérico	*Categorical Number*	**CAT-NM**
	Categórico Codificado	*Categorical Code*	**CAT-CD**
	Categórico Descriptivo	*Categorical Description*	**CAT-DES**
	Categórico de intervalo	*Categorical Interval*	**CAT-IV**
	Categórico Dicotómico	*Categorical Dichotomic*	**CAT-DIC**
	Descriptivo	*Description*	**DS**
	Valor	*Value*	**VAL**
	Valor detallado	*Detail Value*	**VAL-DET**
	Valor agregado	*Aggregate Value*	**VAL-AGG**
	Temporalidad	*Time Measure*	**TM**
	Tiempo - Fecha/Hora	*Date time*	**TM-DTM**
	Tiempo - Fecha	*Date*	**TM-DATE**
	Tiempo - Hora	*Just Time*	**TM-TIME**
	Media binaria	*Binary Media*	**MEDIA**
	Uso no definido	*Not Defined Use*	**NDU**
Origen (*Source*)	Se tiene	*Already Available*	**AA**
	Se calcula	*Calculated Data*	**CAL**
	No disponible	*Not Available*	**NA**
	Origen no definido	*Not Defined Source*	**NDS**
Relacionalidad (*Relationality*)	Atributo primo	*Prime Attribute*	**PK**
	Atributo primo y de llave foránea	*Prime Attribute, Foreign Key Attribute*	**PK-FK**
	Atributo de llave foránea	*Foreign Key Attribute*	**FK**
	Atributo estándar	*Standard Attribute*	**SA**
	Relacionalidad no definida	*Not Defined Relationality*	**NDR**
			V1.2 Desarrollado por: Felipe Ramírez, PhD. 2022

Al clasificar un dato, debemos tomar en cuenta que, con la información disponible, hay tres escenarios de codificación:

a) Sabemos la categoría general, pero no específica (por ejemplo, *QT*);

b) Sabemos la categoría específica (por ejemplo, *QT-DIS* o *QT-DIS-IS*);

c) No tenemos elementos para saber la categoría (por ejemplo, *NDMT*).

Veamos un ejemplo de clasificación. Supongamos que tenemos un dato llamado `distancia`.

Se trata de un número flotante, que al fraccionarse puede ganar precisión y que tiene un cero absoluto (una distancia igual a 0 implica que no hay distancia, o ausencia de magnitud); se trata de un campo con valor detallado, con el cual podremos hacer cálculos, como sumatorias o promedios, aunque no representa en sí mismo el resultado de un cálculo; el valor del dato no lo tenemos, pero sí tenemos un campo llamado `velocidad`, y otro llamado `tiempo`, con los cuales podemos calcular la distancia, siendo entonces un campo calculado. Almacenado en una base de datos, es un atributo estándar, es decir, no forma parte de una llave primaria ni foránea.

Su código sería:

QT-CON-RS/NUM-FL/VAL-DET/CAL/SA

9. Consideraciones generales de la taxonomía

Toma en cuenta que las tipologías no son mutuamente excluyentes entre sí, por ejemplo, un dato puede ser de identificación, numérico, discreto, de escala de intervalo, calculado, todo al mismo tiempo.

Como analista de datos, debes ser capaz de procurar los datos que mejor representen a los fenómenos, y que al mismo tiempo sean lo más compatible posibles con los tratamientos estadísticos y gráficos que le darás a los datos.

Como analista, debes conocer a detalle tus datos, saber su significado, su tipo, sus escalas y su uso.

Ten cuidado de no utilizar un tipo de dato inadecuado para las técnicas estadísticas y gráficas que elijas.