

Fastbook 09

Tratamiento de Datos (Excel y SQL)

Primeros pasos en la validación
de información



09. Primeros pasos en la validación de información

A la hora de realizar un proyecto en el que tengamos que recopilar o utilizar cualquier tipo de información, antes de poder usarlos e integrarlos en nuestros procesos y modelos debemos asegurarnos de que sean coherentes y correctos.

Aunque la validación de estos datos es un proceso muy habitual en cualquier empresa, sigue siendo un aspecto muy abstracto y difuso. De hecho, si pedimos un informe de calidad de la información a distintos profesionales del sector, todos ellos nos devolverán diferentes salidas, y aunque los resultados deberían ser los mismos, pueden estar realizados desde diferentes perspectivas.

Por todo esto, estudiaremos a lo largo de estas páginas los principales aspectos que deberemos tener en cuenta a la hora de realizar un estudio de Data Quality de nuestra información. Además, veremos los distintos enfoques que tendremos que realizar a la hora de analizar nuestra información.

Autor: Breogán Cid

Situación actual

Los pilares de la calidad del dato

Validación de información

Datos numéricos

Textos y variables categóricas

Fechas

Datos geolocalizados

Conclusiones

Situación actual

X Edix Educación

Como ya sabemos, el mundo relacionado con los datos ha ido aumentando su importancia a lo largo de los años, hasta ser considerados como el petróleo del siglo XXI. Este sentimiento ha calado en la mayoría de las empresas y, por miedo a no estar aprovechando las ventajas competitivas que podrían obtener del uso de la información que disponen, lo primero que han realizado es el **almacenamiento masivo de todo dato al que han tenido acceso**, ¿y bajo qué lema?



Yo lo guardo y ya veremos si nos es útil más adelante.

Aunque bien es cierto que las decisiones apoyadas en información son importantes para que las empresas tomen las **decisiones correctas**, tanto a nivel táctico como estratégico, es necesario que los datos usados en los procesos de decisión sean los adecuados, puesto que, de no ser así, los aprendizajes que podríamos obtener podrían convertirse hasta en los opuestos.

Soy consciente de que contado de esta manera parece obvio y que no estoy descubriendo nada que no conozcamos hasta el momento, pero ¿cómo podemos asegurar la calidad de la información?

Muchas veces realizamos nuestras **validaciones automáticamente de forma involuntaria**, por ejemplo, si estamos siguiendo los pasos de una receta de cocina y leemos que tenemos que introducir nuestro pescado al horno 20 minutos a una temperatura de 60°, automáticamente nos daremos cuenta de que la receta es incorrecta, ya que la temperatura indicada no es la habitual en este tipo de recetas.

En otras palabras: nos resultaría extraño a primera vista.

Pero otras veces, estas validaciones no son tan directas e, incluso, será necesario que **el responsable de indicar si la información es correcta o no tenga unos conocimientos mínimos sobre el tema tratado**. Por ejemplo, siguiendo con el caso anterior: si ahora nos indicaran que una vez sazonado nuestro salmón tenemos que dejarlo al horno a 160° en un tiempo de 20 minutos, ¿seríamos capaces de saber si es correcto?

Los pilares de la calidad del dato

X Edix Educación

Debido a la complejidad del proceso de validación de información, no existen estándares en las características o métricas que debemos de tener en cuenta para asegurar que los datos son los adecuados, incluso estas métricas pueden variar de un tipo de proyecto a otros, pero sí que existe **una serie de características o pilares** que todo el mundo acepta y usa, entre los que destacarían:

- La actualización.
- La precisión.
- La integridad.

1

Actualización



Los datos son elementos vivos que cambian constantemente, por lo que es importante que tengamos acceso a la información más reciente posible para que podamos trabajar siempre con el escenario más actual, ya que suele ser el más representativo de la nueva situación.

Esto no quiere decir que tengamos que prescindir de la información histórica, de hecho, en la mayoría de las ocasiones suele ser un elemento clave en la elaboración de modelos o métricas importantes en nuestro negocio, pero todavía hay algo más relevante que no debemos olvidar: **la información actual de hoy se convertirá en los datos históricos del mañana.**

2

Precisión



Como sabemos, **la cantidad de datos que podemos almacenar puede aumentar casi hasta el infinito** si no nos paramos a pensar en nuestras necesidades antes de empezar a almacenarlos.

Otro de los elementos más importantes es la **optimización de nuestras políticas de almacenamiento**, puesto que la presencia de información irrelevante o con mayor granularidad de la necesaria puede hacer que no podamos trabajar con nuestros datos.

Imaginemos, por ejemplo, que estamos trabajando con una empresa del sector retail y necesitamos la **información del clima** para realizar nuestros modelos de ventas a nivel diario. ¿Necesitaríamos disponer de la información del clima a nivel horario? ¿Y de regiones que no estén presentes en el scope del proyecto?

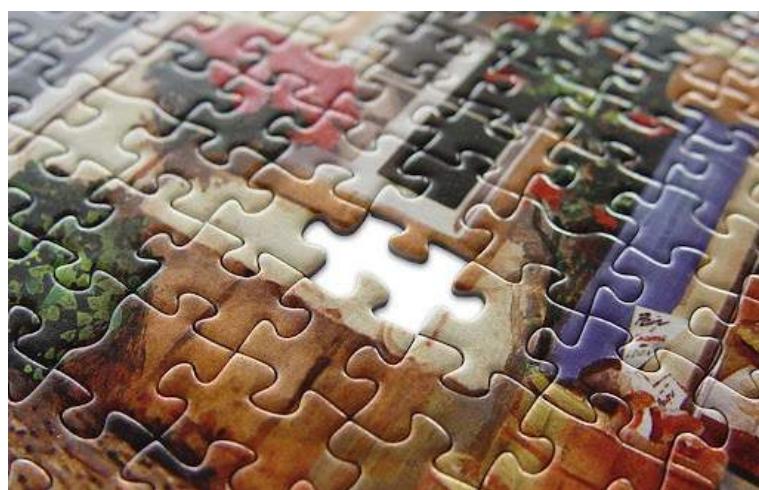
Todas las validaciones se centran en la ausencia de datos o en la tendencia de datos erróneos, pero nos solemos olvidar de la comprobación de la existencia de datos irrelevantes o con mayor granularidad de la necesaria.



Es **importante** que recordemos que los datos nos deberían dar justo lo que necesitamos. En otras palabras: en muchas ocasiones, menos es más. Si conseguimos prescindir de la información irrelevante, no solo minimizaremos las validaciones necesarias, sino que reduciremos las posibles fuentes de errores de nuestro proyecto.

3

Integridad



Uno de los procesos fundamentales en la validación de la información suele ser el contraste de esta en distintas fuentes para confirmar que la información de la que disponemos es la correcta. Este hecho, que a todas luces parece lo más inteligente y correcto, puede ocasionarnos muchos problemas a la hora de trabajar con los datos.

Un claro ejemplo de esto lo hemos vivido durante la pandemia del COVID-19 con la información presente de contagiados y fallecidos. Si nos hemos dedicado a recopilar la información o, al menos, a estar al día de las cifras que se iban produciendo, nos habremos dado cuenta de que esta información variaba en función de la fuente que usáramos. Durante el año 2021 teníamos a nuestra disposición la información actualizada de 5 fuentes distintas de información y rara vez coincidía, a pesar de que toda provenía de la misma fuente oficial.

En nuestro trabajo, tendremos que asegurarnos de que esto no nos ocurre y que no tenemos datos que se contradicen con otros almacenados.

Muchas veces una mala política del almacenamiento de la información hace que los procesos de actualización de información no sean tan eficientes como deberían y que nos dejen inconsistencias en nuestros datos.

Imaginemos el siguiente ejemplo en el que tenemos la información de las ventas para dos productos:

Marca	Fecha	Ventas	Precio
Cocacola	2022-10-01	5055	1.5
Pepsi	2022-10-01	5055	1.5

Tras la actualización de datos, **nuestro cliente analiza la información** se encuentra con un **pequeño error**:

Marca	Fecha	Ventas	Precio
Cocacola	2022-10-01	5055	1.5
Pepsi	2022-10-01	5055	1.5
Coca-Cola	2022-10-01	8053	1.8

Al realizar la actualización, y por problemas en el proceso de carga, se ha duplicado la información por no conservar la clave primaria.

Este ejemplo, por desgracia, es más habitual de lo que podemos esperar, además de ser de las **inconsistencias más complicadas de solucionar**. Para solucionarlo, tendremos que realizar funciones de investigación analizando los ficheros de las fuentes originales para poder eliminar los datos erróneos.

Validación de información

X Edix Educación

Cuando nos surge la necesidad de empezar a validar nuestros datos, tanto si los tenemos disponibles en una base de datos como en los ficheros locales, la primera intención suele ser realizar **las validaciones en la herramienta** en la que nos encontramos, usando consultas en la base de datos o aplicando fórmulas directamente en nuestro Excel, y aunque no quiere decir que sea una mala práctica, siempre que podamos sería conveniente llevarlas a cabo con pequeños scripts realizados en algún lenguaje de programación que conozcamos, ya que, gracias a esto, podremos realizar validaciones automatizadas que pueden ser ejecutadas de forma recurrente tras cada modificación de los datos fuente.

Esta práctica suele conllevar un tiempo extra a la hora de elaborar nuestros informes, pero lo agradeceremos cuando tengamos que validar la nueva información cargada.

No obstante, muchas veces tendremos que **recurrir a validaciones puntuales**, por eso, es importante saber que es tan importante conocer el tipo de validaciones que tenemos que realizar como cuál será el mejor recurso para realizarlas. Sé que puede parecer complicado de entender, por eso, veámoslo a través de un breve ejemplo.

A continuación, nos encontramos con los datos, obtenidos de Aemet, que representan las temperaturas medias de las comunidades autónomas presentes en la Península Ibérica para un día cualquiera. ¿Qué observáis? ¿Qué os llama la atención?

Andalucía	7.491
Aragón	3.254
Asturias	4.905
Cantabria	3.783
Castilla - La Mancha	2.434
Castilla y León	0.422
Cataluña	2.022
Comunidad Valenciana	7.073
Extremadura	5.635
Galicia	5.357
Madrid	1.494
Murcia	8.354
Navarra	1.735
País Vasco	2.924

Para este caso, con una cantidad limitada de datos, podríamos emplear herramientas con las que estemos familiarizados, como Excel, o los lenguajes de programación, como R o Python, y empezar a **buscar valores atípicos**, comparándolos con la media y mediana, o realizar análisis más complejos, pero... La mayoría de las veces siempre se nos pasan las validaciones importantes.



El mayor problema es que nuestro cerebro está habituado a trabajar con imágenes —de ahí el dicho *Una imagen vale más que mil palabras*— y le cuesta mucho trabajar con números.

Si realizamos la **presentación de los resultados anteriores** de una forma más visual, podremos analizar los datos con mucha más facilidad.



Aunque nos hayamos centrado en el análisis de las temperaturas y hayamos usado nuestro conocimiento del territorio para **evaluar si las temperaturas eran las correctas o no**, seguramente se nos habrá pasado uno de los elementos más importantes: los datos no estaban completos, ya que nos faltaba la información de una de las comunidades autónomas.

Es decir, antes de empezar a ver **las validaciones más recomendadas**, según el tipo de datos con los que estemos trabajando, nos tienen que quedar claros **dos conceptos clave**:

- Que las **validaciones son más eficaces** si nos aprovechamos de recursos gráficos que nos permitan transformar un conjunto de datos en imágenes fácilmente reconocibles.
- Que **validar los datos presentes** es lo fácil y lo habitual, pero nosotros tenemos que dar un paso más allá y encontrar la información que falta en nuestros datos.

Ahora que ya hemos entendido estas premisas, empezaremos a ver las distintas consideraciones que debemos tener en cuenta, según el tipo de datos con los que estemos trabajando.

Datos numéricos

X Edix Educación

La validación referente a campos numéricos es **considerada una de las más fáciles de realizar**, ya que, en la mayoría de las veces, se puede apreciar a simple vista. Lo más importante a la hora de empezar a validar esta información es tener claro cuáles son los separadores decimales y de miles que estamos usando, ya que estos pueden cambiar según el idioma en el que tengamos nuestra aplicación.

Para evitar posibles problemas relacionados, se suele aconsejar trabajar siempre con 1 o 2 decimales, pero nunca con 3, ya que visualmente pueden ocasionar muchas confusiones.

Una vez que sabemos que el formato es el correcto, tendremos que aplicar cierto conocimiento específico para validar el contenido. Para ello, no debemos olvidar que un número en sí mismo no nos aporta información si no tenemos en cuenta la unidad a la que va asociada. Por ejemplo, si estamos trabajando con la edad de las personas, debemos tener claro si estamos trabajando en años o en meses (en este último caso si estemos trabajando con información de bebés).

Cuando ya hayamos validado tanto el formato como las unidades, deberemos apreciar si los valores pertenecen al rango esperado de valores permitidos. Para realizar este tipo de validaciones, necesitaremos aplicar el sentido común a nuestros datos, creando los rangos de cada variable según nuestro conocimiento específico y experiencia. Por ejemplo: si queremos validar si la información recogida de la temperatura media de ciertas regiones es correcta, tendremos que conocer a priori cómo es el clima de esa región, ya que puede diferir bastante según su geolocalización en el planeta.

De todas formas, esta formaría parte de las **primeras validaciones** que realicemos sobre nuestros datos, por lo que podemos definir un rango de valores lo suficientemente grande para subsanar la carencia de información relacionada.

Por último, y para poder asegurar la calidad de la información presente, se suelen usar pequeñas comparaciones estadísticas que nos permitan validar la información, relacionándola consigo misma en lugar de validarla contra nuestras hipótesis.

Los métodos más habituales son:

- El análisis de distribución de los valores.
 - La comparación versus su media y su mediana.
-

De momento solo nos quedaremos con esta idea, ya que lo contaremos en detalle en el siguiente fastbook.

Textos y variables categóricas



En muchas ocasiones nos encontraremos en la tesitura de tener que **validar los campos de texto**, pero no os preocupéis, ya que si seguimos los pasos adecuados podréis analizarlos muy fácilmente. Solo debemos tener en cuenta los **siguientes factores**: el uso de mayúscula y minúscula, los espacios en blanco, la codificación, las faltas ortográficas y el idioma.

1

Uso de mayúsculas y minúsculas

Uno de los principales motivos de equivocaciones en las **validaciones de los datos categóricos y en el uso posterior de estos datos** es la creencia de que siempre vamos a trabajar con lenguajes o programas que no son *case sensitive*, o lo que es lo mismo: que no distinguen entre mayúsculas y minúsculas, es decir, para los que ‘Lunes’ es igual a ‘lunes’. Por lo general, cuando vayamos a trabajar con datos de formato texto, sería conveniente, siempre que sea posible, **evitar el uso de mayúsculas**, así nos aseguramos de que la igualdad se cumple en cualquier entorno o lenguaje usado.

2

Espacios en blanco

Otro de los **principales problemas** que nos solemos encontrar en este tipo de validaciones es la presencia de espacios en blanco al inicio y al final de una palabra, o la inclusión de varios espacios en lugar de uno solo, lo que puede hacer que tomemos como diferentes valores a aquellos que deberían ser el mismo, por ejemplo:

- “Miguel Fernández”
- “Miguel Fernández”
- “Miguel Fernández”

En este caso, nos encontramos **distintos elementos** dentro de nuestros datos, pero deberíamos de tratarlos como un único elemento, ¿cierto?

3

Codificación

Como ya sabrás, existen **diferentes codificaciones** utilizadas en nuestro día a día que hacen posible que se puedan crear nuestros documentos en los distintos idiomas usados. Por lo general, y siempre que estemos trabajando en España o en idioma español, lo más habitual es que trabajemos con Latin1 (usado sobre todo en Windows) o con UTF-8 (usado sobre todo en Linux). Normalmente, **no suelen ocasionar problemas**; pero cuando estamos trabajando con caracteres como la ‘ñ’ o con palabras acentuadas, es posible que los mismos caracteres no se representen de la misma manera si estamos trabajando con fuentes de distintos encodings.

Por ejemplo, podemos encontrarnos la palabra España escrita de la siguiente forma: EspaÃ±a.

Son errores poco habituales, pero es fundamental localizarlos
lo antes posible.

4

Faltas ortográficas

Parece mentira, pero uno de los mayores problemas a la hora de analizar estas fuentes de información son las **faltas ortográficas** cometidas por las personas a la hora de almacenar la información. Sobre todo, si estamos trabajando con información que haya podido ser recogida en cuestionarios o campos de texto en cuestionarios. No es la primera vez que me encuentro un ‘Abión’ entre un montón de ‘Aviones’ (perdón, sé que duelen los ojos al ver escrito *avión* con b).

Si bien es cierto que estos errores suelen detectarse con mucha facilidad al trabajar con pocos datos, cuando la cantidad de estos aumenta, **no tenemos la posibilidad de analizar todos los registros de forma independiente**. Entre las posibles soluciones para encontrar estos errores me gustaría destacar dos de ellas.

Agregar los valores distintos y ordenarlos de forma alfabética

Esta solución solo es válida para el caso en que, a pesar de que la cantidad de registros sea numerosa, las distintas opciones del campo sean reducidas (por ejemplo, el género de una persona).

Agregar los valores distintos y ordenarlos por el número de ocurrencias

Esta solución se basa en la idea de que las faltas ortográficas son residuales y que la mayoría de las personas escriben correctamente, por lo que, si analizamos los elementos que casi no se repiten, podremos encontrar la mayor parte de valores erróneos.

5

Idioma

Otro elemento importante es **comprobar el idioma usado** por los usuarios a la hora de crear la información. Aunque la presencia de distintos idiomas en los campos no es un error en sí mismo, tenemos que asegurar que este elemento no se produzca sin nuestro conocimiento. Muchas veces, podremos recurrir a transformar la información a un único idioma, creando una columna adicional de idioma original para no perder la información adicional que nos aporta.

Fechas

X Edix Educación

Los campos de fecha suelen ser uno de los mayores quebraderos de cabeza que nos podremos encontrar si no los estamos tratando bien desde el inicio. Podríamos estar **hablando durante horas y horas sobre estas validaciones**, pero vamos a intentar resumir lo más importante, ya que, como hemos dicho anteriormente, menos es más. Por eso nos centraremos en su formato, validando.

1

Usamos los mismos separadores y los campos se organizan de la misma manera

Existen diferentes formas de escribir la misma fecha y todas son correctas:

- 12-01-2023.
- 01/12/23.
- 12 de enero de 2023.
- 20230112.

Todas estas expresiones hacen referencias a la misma fecha, pero lo más importante es que, a la hora de almacenar nuestros datos, todos los datos se encuentren con el mismo formato.

Además, debemos tener claro a qué hace referencia cada uno de nuestros campos, puesto que ‘12-01-2023’ podría hacer referencia al 12 de enero o al 1 de diciembre según el formato en el que se encuentre la fecha.

Si en nuestros datos tenemos **información de los últimos días de cada mes**, podremos discernir a cuál de las dos opciones hace referencia, pero no siempre es así. Aunque parezca mentira, nos encontramos con muchos problemas en diversos proyectos por estar trabajando con la fecha errónea.

2

Los datos tienen el mismo locale

Tendremos que **validar que los datos corresponden al mismo campo horario**, ya que, si tenemos información de diferentes zonas horarias y no las tenemos unificadas, podremos estar comparando fechas que no son equivalentes.

Como extra, y un elemento para pensar, existen dos días especiales en el año en los que el tiempo cambia, al menos en España... Son los dos días en los que se produce el **cambio horario**. Decimos que son dos días especiales porque si trabajamos con periodicidad horaria o inferior, podremos tener problemas, ya que nos podemos encontrar en el caso de que estemos sobrescribiendo información. Además, también es importante tener en cuenta la existencia de **años bisiestos**, en los que febrero tiene un día más.

Datos geolocalizados

 Edix Educación

Por último, nos centraremos en el tipo de datos que, por lo general, pasa más desapercibido en las validaciones de datos, y (¡atención!) suele ser el **elemento diferenciador entre los buenos y los malos analistas de información**.

Imaginemos que tenemos una **fuente de datos** con la siguiente información:

- ID_cliente.
- Nombre.
- DNI.
- País.
- Provincia.
- Ciudad.
- Latitud.
- Longitud.

¿Qué validaciones creéis que podríamos realizar para asegurar que la información geográfica es correcta?

1

País

Por lo general, siempre que dispongamos del campo 'país', **lo trataremos como la información principal** que nos servirá para validar el resto de información. Tendremos que realizarle las validaciones explicadas para los campos de texto.

2

Provincia

Si nos encontramos con un campo 'provincia', al igual que con el de 'país', tendremos que **realizar las validaciones de los campos de texto**, para identificar, por ejemplo, si existen distintas nomenclaturas para las mismas provincias (véase, A Coruña y La Coruña), pero además, deberemos comprobar que todas las provincias se corresponden con el país indicado previamente.

3

Ciudad

Si nos encontramos con un campo 'ciudad', al igual que con el campo anterior, tendremos que **realizar las validaciones de los campos de texto**, pero además, deberemos comprobar que todas las ciudades se corresponden con la provincia del país indicado previamente.

4

Latitud y longitud

En estos campos nos topamos con la mayor parte de los problemas diarios, ya que muchas veces la imputación de estos se realiza en procesos manuales.

Si disponemos de estos datos, **lo ideal sería poder representarlos en un mapa**, y validar que todos los puntos corresponden con las ciudades indicadas previamente. Gracias a esta validación, podremos detectar varios tipos de fallos:

- Que **los campos vengan invertidos**, lo que puede ocasionar que un punto que nos esperemos en Madrid, por ejemplo, esté situado en medio del océano.
- Que **los datos no sean lo suficientemente precisos** y no coincidan con la ciudad esperada (sobre todo, los puntos cercanos a las fronteras geográficas), o que coincidan con mares o ríos en ciudades costeras.

Este trabajo parece muy complicado, pero **puede hacerse fácilmente mediante el uso de las cartografías gratuitas accesibles en internet en formato SHP**, que pueden ser importadas en la mayoría de los lenguajes de programación.

Conclusiones

X Edix Educación

Como pequeño resumen de lo comentado en este fastbook, quiero **destacar los cuatro aprendizajes fundamentales** sobre el tema que hemos estudiado.

- Hemos visto la importancia de la validación de las fuentes de información, ya que una pronta **detección de errores** evitará que tengamos que repetir trabajo.
- A pesar de la importancia de esta parte del proceso, no hay una manera estándar de realizar los análisis de validación, pero en nuestro caso nos centraremos en tres pilares fundamentales: **la actualización, la precisión y la integridad**.
- También diferenciamos brevemente los distintos tipos de validaciones: las **validaciones recurrentes** que se realizarán sobre todo en los lenguajes de programación para poder ejecutarlas con cada actualización de datos; y las **validaciones puntuales**, en las que intentaremos aprovechar los recursos que conozcamos para poder realizarlas de la forma más ágil posible (por ejemplo, el uso de mapas para analizar información geolocalizada).
- Por último, nos hemos centrado en los distintos aspectos que debemos tener en cuenta en función del tipo de datos que queramos validar.

¡Enhorabuena! Fastbook superado

edix

Creamos Digital Workers