Realiza Calidad de Datos en 4 pasos

Fecha: 13/08/2021

Según Gartner (*Gartner: "5 Steps to Build a Business Case for Continuous Data Quality Assurance", 20 de abril de 2020, Saul Judah, Alan D. Duncan, Melody Chien, Ted Friedman*) "**La mala calidad de los datos destruye el valor empresarial**.

De: Keisa Avila

Para: tí

Guía rápida para aplicar Calidad de Datos.

Un programa de calidad de los datos debe ser una prioridad dentro de una cultura organizacional de data driven, los datos son un activo estratégico que al utilizarlo bien, le da ventajas competitivas a las organizaciones.

En mis proyectos, luego que realizó la extracción de la data, tengo por costumbre aplicar calidad de datos a la información y me baso estos cuatro pasos:

1.- Análisis de Nulos y Atípicos

1.1.- Nulos

En cualquier conjunto de datos, cuando le aplicamos la calidad de datos, nos vamos a enfrentar con datos nulos, que significa datos que no vienen o no existen.

Imagínate que tiene un conjunto de datos de 1,000 registros y hay 100 registros que le falta una variable.

Ejemplo:

El campo de **fecha de alta** del cliente, es importante para el negocio y como regla es necesario que ese dato no esté vacío.

En un conjunto de datos de 1,000 registros, existen 900 registros que sí poseen ese dato de fecha de alta y 100 están nulos.

Entonces se puede decir que esa variable de **fecha de alta**, con respecto al análisis de nulos tiene un 90% de calidad de dato, porque el otro 10% están nulos y hay que tomar acciones para arreglarlo.

El análisis de datos nulos no es tan fácil determinarlo y solo contar ¿Cuántos registros vienen con una variable llena? o ¿Cuántos registros me vienen con una variable vacía?, porque ese dato puede que sea nulo por naturaleza y según la estructura de información del negocio ese dato esta bien. Para realizar este tipo de análisis, tenemos que tener conocimiento de ¿Cuáles son las reglas del negocio? y estar claro si ese dato representa un problema o está bien por su naturaleza.

1.1.- Atípicos

Un valor **atípico** es un dato que es considerablemente diferente a los otros **datos** del conjunto, son datos raros pero pueden ser reales.

Ejemplo:

El ejemplo más común es la edad, puede ser que tengamos en nuestro conjunto de datos un cliente que tenga 98 años, no es un dato de los más frecuentes pero no quiere decir que sea un error entonces lo clasificaremos como dato atípico, en cambio si conseguimos un dato con 180 años eso si es un error.

2.- Estadísticas Básicas

En este paso sacamos las estadísticas básicas del conjunto de datos con respecto a una variable.

- ☐ El mínimo
- ☐ El máximo
- ☐ El promedio
- ☐ La media
- ☐ La mediana
- ☐ Si tiene valores negativos o no los tiene
- ☐ Rango de la variable

Con estas estadísticas básicas, se detectan los problemas de calidad, si seguimos con el ejemplo de la edad, cuando la máxima me arroja un valor de 180 años ya sabemos que hay un error o la parte de valores negativos hay variables que no pueden ser negativa por ejemplo de años de graduado, ese dato nunca va a ser negativo.

3.- Análisis Longitudinal

Se identifican las variables y se recogen los datos cualitativos y cuantitativos en un periodo de tiempo, esto se realiza cuando hay relaciones prolongadas o continuas entre el negocio y el cliente.

Ejemplo:

Un banco cuando le concede un préstamo de auto, en ese caso comienza una relación prolongada, lo normal es que en los siguientes meses, el cliente mantenga unos datos prolongados, regulares y muy parecidos a través del tiempo.

Ejemplo del préstamo de auto:

Mes	Deuda $	Calidad del dato
Enero	$ 15,000	Sin error
Febrero	$ 14,900	Sin error
Marzo	**$ 3,000**	**Con error**
Abril	$ 14,700	Sin error

En este caso si analizamos la variable sola, no genera ningún tipo de error, porque un cliente puede deber 3,000 dólares en marzo, pero si revisamos la variable a través del tiempo nos damos cuenta de que existe un error.

4.- Coherencia entre variables.

Es la relación lógica entre 2 variables, pasa igual que el análisis longitudinal, la variable sola no nos dice que tiene error, pero si le busca la lógica con otra variable allí si consigues el error.

Por ejemplo:

Cantidad de productos comprados: 0

Monto en productos comprados: 500 $

En este caso no se sabe, ¿Cuál de las dos variables está mala?, pero el caso es que si las analizas juntas, una compra por 500$, con una cantidad de productos en 0, es un error.

Estos son los 4 tipos de calidad de datos, que yo aplico de forma rápida a mis proyectos, para tener una idea general de como esta la data. Espero que te sirva de ayuda y lo utilices como una guía para comenzar a realizar los análisis. Es recomendable que antes de presentar un dashboard al negocio para la toma de decisiones realices de manera rápida, un análisis de calidad de datos, para detectar los errores. Existen análisis más profundos, pero te quería dar mis cuatro pasos.