Preparación de los datos – Integridad de los datos

En este video quiero hablar sobre la integridad de los datos y por qué es realmente importante conseguirla. Esto será muy importante mientras hacemos análisis de datos o mientras limpiamos datos.

Quiero hablar sobre cosas como esta: tal vez están hablando con un funcionario del gobierno y le están pidiendo una tabla para hacer un análisis computacional más tarde, y luego obtienen algo como esto de aquí o algo así. Y quiero hablar sobre las diferencias entre estos archivos, cosas que vemos aquí que hacen que esta no sea la tabla que queremos, y que hacen que estos datos no estén estructurados de una manera que nos sea útil en nuestro flujo de trabajo de periodismo de datos.

En estas tablas aquí pueden ver que hay elementos que agregan los datos. Quizás aquí tenemos el total, tenemos subcategorías, tenemos espacios vacíos, tenemos colores. Estos son todos los elementos que hacen que esta tabla sea agradable para el ojo humano, ¿verdad? Aquí sucede lo mismo: hay espacios vacíos para facilitar la identificación de que todo aquí está conectado con esto, y luego tenemos un encabezado aquí en un color diferente, y tenemos columnas aquí que se fusionan, y tenemos la suma.

Estas son todas las tablas que se agregan para que las podamos analizar. Pero, de hecho, estas tablas están destinadas a ser leídas en una pantalla o impresas en papel, son tablas para humanos. Por lo general, presentan datos consolidados que provienen de tablas estructuradas, y estas son las tablas que queremos, con los datos granulares que generaron estos reportes. La cuestión es que estos reportes que estamos viendo aquí generalmente son tablas dinámicas e informes que son archivos PDF, son inútiles para el análisis computacional porque no tienen integridad en términos de previsibilidad para hacer cálculos, y hablaré sobre por qué esas cosas son diferentes.

Entonces, cuando miramos cosas como estas, esto presenta los meses en portugués y también tenemos datos de seguridad en Brasil. Tenemos el año aquí y tenemos un tipo diferente de dato aquí. Tenemos “n/d” aquí, que puede ser un “no disponible” o “sin datos”. Pero este es muy diferente de los informes que presenté anteriormente. O este, que presenta el cuadro de medallas de los primeros Juegos Olímpicos de la era moderna. Esto es oro, plata y bronce, todo en portugués. Tenemos aquí el año, la ciudad y el país, y luego el tipo de medalla. Estas son tablas hechas para máquinas, ¿verdad? Son legibles por máquina. Pueden ser procesadas por software especializados y nos permiten crear tablas para humanos, como mencionamos anteriormente. Y son excelentes para el análisis computacional.

Y déjenme hablar sobre las diferencias entre estos dos tipos de tablas que coexisten en el mundo. Aquí está la tabla estructurada que mostré sobre los datos de seguridad en Brasil. Y esto puede sonar muy básico, pero es muy importante cuando buscamos conjuntos de datos y limpiamos y analizamos conjuntos de datos. Entonces, por ejemplo, en una tabla estructurada,

generalmente tenemos filas aquí en azul y también tenemos columnas. En cada una de esas columnas y filas tenemos solo un tipo de dato. No tenemos dos tipos de datos. No tenemos números y luego texto, o nombres de ciudades y países en la misma columna. Solo tenemos un tipo de dato. También tenemos aquí en azul lo que llamamos la clave primaria, el número de fila que está asociado con un registro dado. Y aquí tenemos el nombre de las columnas, que nos ayudará a identificar qué hay dentro de la columna para que podamos hacer cálculos más adelante. Aquí vemos, por ejemplo, en la columna del año, tenemos el año 1896 y en la columna del país tenemos aquí “Estados Unidos”, en portugués. Entonces vemos que solo hay países aquí, solo ciudades y solo años.

Ahora echen un vistazo a este. Este es un conjunto de datos, una tabla que muestra los aeropuertos, diferentes aeropuertos en Brasil, pero también muestra la abreviatura de los estados junto al nombre del aeropuerto. También muestra el total aquí en la parte superior. Y suma y mezcla números con porcentajes. Esto muestra la cantidad de vuelos que se han retrasado o cancelado en Brasil en el pasado reciente. Entonces, aquí en la primera columna, generalmente la columna, como mencioné, tiene el nombre de los datos, un nombre que describirá lo que está sucediendo aquí en esa columna. Pero solo tenemos un tipo de dato. Debemos tener solo un tipo de dato por columna. Y aquí comenzamos con un total que no debería estar aquí. También está el nombre de la ciudad y el acrónimo del estado. Por lo tanto, no deberíamos tener estos dos datos, estos dos tipos de datos coexistiendo en la misma columna. Necesitamos dividirlos para que tengamos dos columnas, una que se ocupe de las ciudades y otra que nos muestre los acrónimos de los estados, tal vez. Y aquí lo mismo. Tenemos el número absoluto de vuelos que se retrasaron y los vuelos que se cancelaron, y luego tenemos el porcentaje de vuelos que se retrasaron y cancelaron. Nuevamente, necesitamos separar estos dos números, separarlos en dos columnas diferentes, de modo que solo tengamos una columna para cada tipo de dato.

Esto será útil porque si necesitamos hacer un cálculo, por ejemplo, una suma, o una división, o promedio, o el número más alto, no tendremos nada más que compita con los datos dentro de la columna. Por lo tanto, es realmente importante tener un solo tipo de dato para que podamos hacer este tipo de cálculos. Entonces, como mencioné, cada columna tiene solo un tipo de dato y, por supuesto, se deben eliminar las filas duplicadas, de lo contrario, agregarán valores que no deberían agregarse.

Pero también tengan cuidado con los errores tipográficos engañosos, ¿verdad? “puerta” sigue siendo diferente de “Puerta” con P mayúscula. Estas son dos entidades diferentes para un computador. Aunque un ser humano puede no reconocer la diferencia entre “puerta” y “Puerta” y puede pensar que no es gran cosa, estas son entidades totalmente diferentes para un computador. Por lo tanto, es muy importante mantener el mismo tipo de escritura para todos los valores que tenemos en nuestro conjunto de datos. Podemos hacer todo en mayúsculas o todo en minúsculas, pero necesitamos un valor estándar que apliquemos a todos los valores. Lo mismo para este, que tiene un espacio vacío antes de “ Puerta”, y los espacios vacíos que están al lado o después de las palabras, antes de las palabras, también se cuentan como

entidades diferentes en una columna. Por lo tanto, debemos asegurarnos de tener una escritura estándar para todos los valores del conjunto de datos. Por eso es importante, en la fase de limpieza, eliminar todos los espacios o corregir errores tipográficos para que los datos en la columna tengan integridad.

Y como mencioné también, la primera fila generalmente tiene el nombre de las columnas y generalmente evitamos incluir datos agregados como los totales en la fuente de datos sin procesar. Esto se debe a que vamos a utilizar un software que procesará los datos, por ejemplo, aplicará fórmulas de sumatoria, por lo que no necesitamos que se agregue el total o que estén encima de los datos, los datos granulares, los datos sin procesar que tenemos. Entonces, lo que generalmente hacemos es eliminar todos los totales, y les mostraré ejemplos en la clase de limpieza de datos sobre cómo podemos asegurarnos de que estos valores se eliminen para que tengamos un conjunto de datos limpio para hacer cálculos más adelante.

Y ¿cuál es el formato más común? Por lo general, cuando hablamos de bases de datos, estamos hablando de archivos CSV la mayor parte del tiempo. O queremos tener un archivo CSV porque es un formato abierto, es texto sin formato, es compatible con casi todas, si no todas las aplicaciones de procesamiento de datos. Y CSV significa valores separados por comas o incluso valores separados por caracteres.

Para crear un archivo CSV – también se le llama TSV si es un separador diferente – pueden crearlo con mucha facilidad. Un archivo CSV son solo valores que están separados por un delimitador. Y el delimitador puede ser una coma, o puede ser un punto y coma, o puede ser este carácter llamado barra vertical, o incluso puede ser Tab. Cuando presionan Tab y este les da este espacio, incluso puede ser ese el carácter cuando crean su CSV o si es un tab, normalmente querrán nombrarlo un archivo TSV.

Entonces, simplemente escriben, pueden crear, ya saben, en un editor de texto, simplemente escriban los nombres y luego pongan una coma. Y, en general, lo que sucede es esta primera fila aquí, estas serán las columnas de la fila y cuando creen una nueva fila, la nueva fila será la nueva fila en nuestro conjunto de datos, en nuestra base de datos. Y luego tenemos el valor aquí, el primer valor será para la primera columna, el segundo para la segunda, y así sucesivamente.

Entonces, miren lo que estamos haciendo aquí con esto es crear algo como esto. Este código aquí, el código CSV aquí, se creará de hecho cuando lo veamos en un programa de hoja de cálculo como Google Sheets, de hecho verá la primera fila aquí y luego la siguiente fila aquí con los valores para cada una de las columnas. Es por eso que es tan importante que los archivos CSV tengan la misma cantidad de elementos en cada línea, para que los datos no se vuelvan confusos, ¿correcto? Entonces si tienen un archivo que tiene siete columnas como este, la segunda línea también debe tener siete valores. Incluso si es un valor vacío, necesita una coma que indique que es un valor vacío.

Los delimitadores más comunes son – como mencioné – coma, punto y coma, barra vertical y también Tab, cuando presionan Tab en su teclado. Entonces, cuando crean archivos CSV, afortunadamente no tienen que escribir todos los valores ustedes mismos en un editor de texto. La mayoría de las aplicaciones nos exportarán archivos CSV. Entonces, solo presten atención al delimitador que se utiliza para generar el archivo CSV. A veces intentan importar un archivo CSV y luego algo sale mal, no reconoce el delimitador y, según el país en el que se encuentren, existe una convención sobre el tipo de delimitador que utilizan. Por ejemplo, en Estados Unidos es una coma. Pero en Brasil, el valor estándar es el punto y coma. Esto crea cierta confusión a veces, por lo que es posible que también deseen verificar esto si no funciona en sus casos.