Preparación de los datos – Limpiando datos con Google Sheets

En este video, les mostraré cómo usar algunas funciones y recursos de Google Sheets para limpiar su conjunto de datos y mejorarlo para su análisis y sus visualizaciones.

En este video usaré el conjunto de datos que creamos al raspar el sitio de Billboard con las 200 canciones principales. Y lo primero que notarán es que ordené las canciones aquí y vine aquí en la columna E y seleccioné “Ordenar página A-Z”. Y se dan cuenta de que cometí un error, importé este conjunto de datos dos veces, esto puede suceder. Así les daré una idea de cómo eliminar duplicados, ¿correcto? Así que ahora tengo muchos duplicados y necesito eliminar todos estos duplicados. Entonces, ¿qué debo hacer?

Google Sheets tiene un recurso realmente genial aquí en “Datos” (Data), puede seleccionar “Eliminar duplicados” (Remove duplicates). Cuando seleccionen esto, diré que mi conjunto de datos tiene una fila de encabezado y luego haré clic en “Eliminar duplicados”. Luego este encuentra los 200 duplicados y los elimina. Muy fácil.

OK, ahora también podrían querer eliminar estas pequeñas columnas aquí. Entonces, solo seleccionaré ambas, hago clic aquí en la letra A y luego mantengo presionada la tecla Shift, y luego hago clic aquí también en B, luego hago clic derecho y luego elimino las columnas A y B. Listo. Está mucho más limpio ahora.

Ahora notarán que las direcciones de las imágenes no están completas para algunas de ellas, ¿verdad? Aquí tenemos chart-static.billboard.com. Esta parece que está completa, pero hay una doble barra diagonal aquí al principio. Vean que otras no las tienen, esto es para imágenes diferentes que fueron capturadas. Así que queremos asegurarnos de que todo esto aquí permanezca igual. O reemplazamos todos los “https://” con las dobles barras diagonales, o reemplazamos las barras con “https:”. Entonces, lo que voy a hacer es reemplazarlo todo aquí porque quiero la dirección completa. Reemplazaré la barra doble diagonal con “https://”, y para eso seleccionaré esta columna y luego digitaré “Command+F” en un Mac, o CTRL, Command+F en un Mac y CTRL+F en un Windows o Linux. También pueden ir a Editar y Buscar y reemplazar.

Si hacen clic aquí, abren esta ventana y lo que queremos hacer aquí es que hay una parte específica que ya está seleccionada aquí y queremos reemplazar solo las barras diagonales aquí. No estas barras diagonales, porque si escribimos aquí “buscar todas las barras y reemplazarlas con “https://”, si hacemos eso, veamos qué sucede. Reemplazará todos, pero luego los que ya existen también se reemplazan también por lo que duplicamos todo lo que está sucediendo aquí. Entonces no queremos eso, ¿verdad? Lo que queremos es una forma de reemplazar solo las barras diagonales que comienzan aquí en la fila. Así que vayamos de nuevo aquí, “Editar”, “Buscar y reemplazar” y usemos algo que también usamos para Web Scraper, que son expresiones regulares, pero eso es bastante fácil.

Queremos indicar que queremos todas las barras al comienzo de la línea y lo hacemos escribiendo esto en “Buscar”. Así que escribimos esto y luego “//”, y nos aseguramos de tener la casilla “Buscar usando expresiones regulares” seleccionada. Vamos a reemplazar esto con “https://” y luego reemplazarlo todo. Y sí, ahora tenemos todas las filas aquí con las direcciones correctas y estamos listos para seguir adelante.

Recuerden que mencioné que capturamos la fecha, pero la fecha aquí no es muy útil en este momento porque hay mucha información aquí. Sí, es una fecha, pero no está en un formato que podamos usar o que nos sea útil. Podemos convertir esto en un formato de fecha completo o podemos separar el mes, el día y el año. Ahora les voy a mostrar algunas técnicas porque pueden hacerlo a través de diferentes técnicas, separar estos valores. Les mostraré algunas pocas de ellas.

La primera es solo extraer el año, y el año que viene de la derecha, es solo uno, dos, tres, cuatro caracteres desde la derecha. Hay una fórmula llamada “right” (derecha) en Google Sheets. Simplemente pueden escribir “right” y luego seleccionan la cadena de caracteres de la que desean seleccionar y luego seleccionarán el número de caracteres que desde la derecha seleccionarán. En este caso son cuatro. Es uno, desde la derecha, uno, dos, tres, cuatro. Y luego hacen esto, y obtienen sólo el año. Y si hacen doble clic aquí, en esta esquina inferior derecha de la celda, hagan doble clic aquí y aplican la fórmula a todas las celdas que están abajo. Entonces esto es solo para extraer el año.

Ahora, si quisiéramos extraer el mes, podríamos usar otra fórmula llamada “left” (izquierda). Así, con esta fórmula, escribimos “left” y luego comenzamos a contar desde la izquierda, que son uno, dos, tres, cuatro, cinco, seis caracteres. Entonces desde la izquierda aquí queremos esto, y queremos seis caracteres. Y aquí es agosto. OK. Hacemos doble clic aquí en la esquina inferior derecha y también lo tenemos todos los meses abajo. Listo.

¿Qué pasa si queremos el día aquí en el medio? Este será un enfoque mixto porque hay una fórmula, algunos de ustedes habrán adivinado, llamada “mid” para extraer lo que está en el medio. Pero para eso primero necesitamos información. Necesitamos decirle a la función “mid”, lo pondré aquí “mid”, necesitamos decirle a la función “mid” qué es esta cadena. Pero desde qué posición, ¿correcto? Entonces necesitamos decir cuál es la posición de “31” aquí. Entonces es uno, dos, tres, cuatro, cinco, seis, siete, ocho. Entonces comenzamos en 8, y la longitud exacta es 2, y luego obtenemos 31. Bien, entonces aplicamos eso a todo.

Ahora hay otra función que lo hace todo más fácil, que se llama “Split”. Split (dividir). Lo que hace “Split” es tomar una cadena, que es esta, y luego le dan un separador o delimitador, e intentan romper la cadena usando este delimitador. Aquí en “Agosto 31, 2019”, observen que hay un espacio entre todas las palabras, por lo que podemos usar este espacio como delimitador para dividir la cadena en diferentes valores. Así que lo abriré aquí entre comillas, pondré un espacio y veré qué sucede. Y vean que la función de división divide

automáticamente este valor aquí en tres valores separados e incluso ignora la coma aquí, por lo que no tienen que lidiar con eso.

Pero hay una manera aún más fácil de hacer esta división. Copiaré la columna aquí y les mostraré este nuevo recurso. Pueden venir aquí debajo de “Dividir texto en columnas” (Split text in columns) y convertirá automáticamente los valores en columnas y detectará automáticamente el separador aquí.

Así que aquí tienen muchos enfoques para extraer información de este valor y convertirla en propia. En este caso aquí tenemos el año, y luego el mes y el día. Eliminaré esto, y ahora tenemos un conjunto de datos mucho más limpio para trabajar, porque ahora tenemos más información. Tenemos información sobre la semana, el año. Incluso podemos eliminar esto, esta columna aquí. Pueden eliminar esto, y ahora tenemos el conjunto de datos aquí.

Por lo tanto, estas son solo algunas estrategias para limpiar su conjunto de datos y mejorar su análisis o visualización posterior. Entonces, vayan a Google Sheets e inténtenlo.