Preparando os dados - Limpando os dados com Google Sheets

Neste vídeo eu vou mostrar como usar algumas funções e recursos do Google Sheets para limpar seu conjunto de dados e torná-lo melhor para sua análise e suas visualizações.

Neste vídeo eu vou usar o conjunto de dados que nós construímos ao raspar o site da Billboard e sua parada com as 200 canções mais tocadas. E a primeira coisa que você vai notar é que eu ordenei aqui as canções e vim aqui na coluna E e selecionei "Classificar página A-Z". E você percebe que eu cometi um erro, eu importei este conjunto de dados duas vezes, para que isso acontecesse, para mostrar como remover duplicatas, certo? Então eu tenho um monte de duplicatas agora e eu preciso remover todas estas duplicatas. Então, o que eu devo fazer?

O Google Sheets tem um recurso muito legal aqui em "Dados", você pode apenas selecionar "Remover cópias". Quando você seleciona isso, eu vou dizer que o meu conjunto de dados tem uma linha de cabeçalho e depois vou apenas clicar em "Remover cópias". Ele então encontra todas as 200 duplicatas e as remove. Muito fácil.

Ok, então agora eu poderia querer excluir estas colunas aqui também. Então, eu vou apenas selecionar as duas, eu clico aqui na letra A e então eu seguro a tecla Shift, e então eu clico aqui na B também, depois eu clico com o botão direito, e então eu apago as colunas A e B. Tudo certo. Está muito mais limpo agora.

Agora você vai notar que os endereços para as imagens não estão completos para algumas delas, certo? Temos aqui charts-static.billboard.com. Esse parece que está completo, mas tem essa barra dupla aqui no começo. Você vê que outros não têm, são para imagens diferentes que foram capturadas. Então nós queremos ter certeza de que tudo isso aqui permaneça igual. Ou nós substituímos todos os "https://", ou substituímos as barras com "https:". Então o que eu vou fazer é substituir tudo isso aqui porque eu quero o endereço completo. Vou substituir a barra dupla por "https://", e para isso eu vou selecionar esta coluna e então eu vou digitar Command + F em um Mac, ou um ctrl, Command + F em um Mac e ctrl + F no Windows ou Linux. Você também pode ir em Editar e Localizar e substituir.

Se você clicar aqui, você abre esta janela e o que queremos fazer aqui é que há uma parte específica que já está selecionada aqui e queremos substituir apenas as barras aqui. Não essas barras, porque se a gente digitar aqui "encontre todas as barras e substitua por https://", se fizermos isso, veja o que acontece. Ele substitui todas elas, mas então as que já existiam são substituídas por isso também, então duplicamos tudo que está acontecendo aqui. Então nós não queremos isso, certo? O que queremos é uma maneira de substituir apenas as barras que começam aqui na linha. Então vamos aqui novamente, "Editar", "Localizar e substituir" e vamos usar algo que usamos para o Web Scraper também, que são as expressões regulares, mas isso é bem fácil.

Queremos indicar que queremos todas as barras no começo da linha e fazemos isso escrevendo isso em "Encontrar". Então digitamos isso e depois "//", e certifique-se de que você tem a caixa "Pesquisar usando expressões regulares" selecionada aqui. Vamos

substituir isto por "https://" e então substituímos tudo. E sim, agora temos todas as linhas aqui com os endereços corretos e estamos prontos para avançar.

Então lembre que eu mencionei que nós capturamos a data, mas a data aqui não é muito útil neste momento porque tem muitas informações aqui. Sim, é uma data, mas não está em um formato que possamos usar ou que seja útil para nós. Podemos transformar isso em um formato de data completo ou podemos separar o mês, o dia e o ano. Agora eu vou mostrar algumas técnicas porque você pode fazer isso por meio de técnicas diferentes para separar estes valores. Vou mostrar algumas delas.

O primeiro é apenas para extrair o ano, e o ano vindo da direita, ele é apenas um, dois, três, quatro caracteres da direita. Há uma fórmula chamada "right" ("direita") no Google Sheets. Você pode apenas digitar "right" e então você seleciona a cadeia de caracteres da qual você quer selecionar e, em seguida, você vai selecionar o número de caracteres da direita que você vai selecionar. Neste caso são quatro. É um, a partir da direita, um, dois, três, quatro. E depois você faz isso, e então você obtém apenas o ano. E se der um clique duplo aqui, neste canto inferior direito da célula, clique duas vezes aqui e você aplica a fórmula a todas as células que estão abaixo. Então isso é apenas para extrair o ano.

Agora, se quiséssemos extrair o mês, podemos usar outra fórmula chamada "left" ("esquerda"). Assim, com essa fórmula, nós digitamos "left" e então começamos a contar a partir da esquerda que é um, dois, três, quatro, cinco, seis caracteres. Então, a partir da esquerda aqui queremos isso, e queremos seis caracteres. E aqui está agosto. OK. Nós clicamos duas vezes aqui no canto inferior direito e também temos todos os meses abaixo. Tudo certo.

E se quisermos o dia aqui no meio? Essa vai ser uma abordagem mista porque existe uma fórmula, alguns de vocês podem ter adivinhado, chamada "mid" para extrair o que está no meio. Mas para isso precisamos primeiro de algumas informações. Precisamos dizer à função "mid", vou colocar aqui, "mid", precisamos dizer à função "mid" o que é essa cadeia de caracteres. Mas a partir de que posição, certo? Então precisamos dizer qual é a posição do "31" aqui. Então é um, dois, três, quatro, cinco, seis, sete, oito. Então começamos no 8, e o comprimento exato que é 2, e em seguida obtemos 31. Ok, então aplicamos isso em tudo.

Agora, há outra função que faz tudo isso de maneira mais fácil, que é chamada de split, dividir. O que o split faz é pegar uma cadeia de caracteres, que é esta aqui, e então você dá um separador ou um delimitador, e ele tenta quebrar a cadeia usando este delimitador. Aqui em "August 31, 2019", perceba que há um espaço entre todas as palavras, então podemos usar este espaço como o delimitador para dividir a cadeia de caracteres em valores diferentes. Então eu vou abrir aqui entre aspas, vou colocar um espaço, e vou ver o que acontece. E veja que automaticamente a função split divide este valor aqui em três valores distintos e até ignora a vírgula aqui, para que você não tenha que lidar com isso.

Mas há uma maneira ainda mais fácil para fazer essa divisão. Eu vou copiar aqui a coluna e vou mostrar este novo recurso. Você pode vir aqui em "Dividir texto em colunas" e ele vai

automaticamente converter os valores em colunas e você vai automaticamente detectar o separador aqui.

Então aqui você tem muitas abordagens para extrair informações deste valor e transformá-lo em seu. Neste caso aqui temos o ano, e então o mês, e o dia. Vou deletar isso, e agora temos um conjunto de dados muito mais limpo para trabalhar, porque agora temos mais informações. Temos informações sobre a semana, o ano. Podemos até apagar isso, esta coluna aqui. Pode excluir esta, e agora temos aqui o conjunto de dados.

Então, essas são apenas algumas estratégias para limpar seu conjunto de dados e torná-lo melhor para analisar ou visualizar depois. Então, vá ao Google Sheets e experimente.