Preparing Data - Cleaning data with Google Sheets

In this video I'm going to show you how to use some Google Sheets functions and features to clean your data set and make it better for your analysis and visualizations.

In this video I'm going to be using the data set that we have built out of scraping the Billboard's website with the top 200 songs. And the first thing that you'll notice here is that I ordered here the songs so I went here to column E. And then I selected "Short sheet A-Z". And you notice that I made a mistake, I imported this data set twice, so this might happen. To give you a sense of how to remove duplicates. Right. So I have a lot of duplicates now and I need to remove all of these duplicates. So what should I do? Google Sheets has a very cool feature here in "Data", you can just select "Remove duplicates". So when you select this, I'm gonna say that my dataset has a header row and then I'm just going to click on "Remove duplicates". Then it finds all of the 200 duplicates and removes them. Very easy.

Ok, so now I might want to delete the smaller columns here too. So I'm just going to select both of them, I just click here on the letter A and then I hold shift, and then I click here on B too, then I click right click, and then I delete columns A and B. All right. So it's much cleaner now.

Now you will notice that the addresses for the images, they are not complete for some of them. Right. We have here charts-static.billboard.com. This looks like it's complete but it has this double slash here at the beginning. You see that others don't have. So that's four different images that were captured. So we want to, we want to make sure that all of this stays the same. Either we replace all of the "https://", double forward slashes or we replace the four slashes with "https:".

So what I'm going to do is to replace all of this because I want the full address. I'm going to replace the double forward slash with "https://", and to do this I'm going to select this column and then I'm going to go and hit "command+5" on a Mac, or a control, "command+F" on a Mac and "control+F" on a Windows or Linux machine. You can also do edit, and find and replace. If you click here, you open this window and what we want to do here is that there is a specific range that it's already selected here and we want to replace only the forward slashes here. Not these forward slashes, because so if we type here "find all the forward, the double forward slashes and replace with "https://", if we do this look what happens. It replaces all of them, but then the ones that already exist they get replaced with this too so we duplicate everything that's going on here. So we, we don't want that.

Right. So what we want is a way to replace only the forward slashes that begin here on the row. So we're gonna go here again, "edit", "find and replace" and we're gonna use something that we use for the web scraper too which are the regular expressions but, it's fairly easy. So we want to indicate that we want all the four slashes on the beginning of the line and we do this by typing this on the "Find". So, we type this and then, "//" and then make sure that you have the "search using regular expressions" box here marked. We're going to replace this with "https://" and then we replace all. And yeah now we have all of the rows here with the correct addresses and we're ready to move forward.

So remember that I mentioned that we captured the date, but the date here it's not very useful at this moment because it has lots of informations here. Yes it is a date but it's not in a format that we can use or that is useful for us. We can either transform this into a full date format or we can split the month, the day and the year. And now I'm going to show you a few techniques because you can do this by different techniques to split these values. So I'm going to show you a few of them.

The first one it's just to extract a year, and the year coming from the right, they're just like one, two, three, four, characters from the right. So there is a formula called "right" in Google Sheets. You can just type "right" and then you select the string that you want to select from and then you're going to select the number of characters that from the right you're going to select. So in this case it's four. It's one, from the right, one, two, three, four. And then you do this, and then you get only the year. And if double click here, on this lower right corner of the cell, you double click here. Then you apply the formula to all of the cells underneath it. So this is just to extract the year.

Now if we wanted to extract the month we can use another formula called "left". So with left, we type "left" and then we're gonna start counting from the left which is one, two, three, four, five, six, characters. So from the left here we want this, and we want six characters. And this is August. Okay. We double click here on the right lower corner and we also have all of the months below. All right.

So, what if we want the day here in the middle? We have, this is going to be a mixed approach because there is a formula, some of you might have guessed, called "mid" to extract what's in the middle. But to do that we need information first. We need to tell the "mid" function. I'm going to put here "mid". We need to tell the mid function what is this string. But then, starting at what position. Right. So we need to tell what is the position of the "31st" here. So it's like one, two, three, four, five, six, seven, eight. So we start at "8", and then the exact length which is "2", and then we get 31. Ok, so we apply it for everything.

Now, there is another another function that makes all of this way easier which is called split, "SPLIT". So what split does is that it takes a string, which is this one, and then you give a separator or a delimiter, and then it tries to break down the string by using this delimiter. So here in "August 31st, 2019" notice that there is a space between all of the words so we might use this space as the delimiter to split the string into different values.

So I'm just going to open here like between quotes, I'm just going to put a space, then I'm going to see what happens. And see that automatically the split function splits this value here in three separate values and even ignores the comma here, so that you don't have to deal with that but there is even an easier way, to do this split. So I'm just going to copy here the column and then I'm going to show you that it's a, it's a new feature. You can come here in "Split text to columns" and it will automatically convert the the values into columns and you will detect automatically the separator here. So here you have like lots of approaches to extract information from this value and make it your own.

So in this case here we have the year, and then we have, and then we have the day. And then delete this. And now we have a much cleaner data set to work on because now we have more information, we have information about the week, the year, we can even delete this one. This column over here. Can delete this one, and now we have here the data set.

So, just to feel a few strategies to clean your data set and make it better to analyze or visualize at a later stage. So, go to Google Sheets and give it a try.