Preparing Data - Cleaning with Google Cloud Dataprep

I'm gonna show you how to use some techniques to clean a data set that's very dirty. I'm gonna be using a great example that was put together by Sarah Cohen. And she was still at the New York Times, she's now at the Arizona State University. It was a great example. It was a class that I watched her give during the computer assisted conference, computer assisted reporting conference in 2016 in Denver.

It is a great example because this is a data set that has everything in it. It has typos, it has blank spaces, it has everything. So, this is a data set that is compiled from Medicaid long-term managed care reports from the New York State. And it looks like this. So, it's very messy, it has on one place here like the name of the plans. You have the counties in New York State and then have the total enrollment. Here you have kind of the date of the month and then the year. But then you have different patterns for that, and it really is messy and cheap. Every sheet on the spreadsheet was for a month in a given year and then she built like a full spreadsheet only one sheet with all of the information that we need, right?

So we want to come from this, to something like this. It's much tidier. So we have the plan names in one column, we have the county in another one, we have the number of enrollments, the month and the year. So each variable has its own column, every column has only one type of data and we have everything together here. We also are eliminating like the totals, right? Because we don't need those. We can calculate later so we don't have, we don't need something here that will tell us about total because we're gonna be using the processing application to do that. So the application that I'm gonna be using today, so she gave a class on OpenRefine and I'm going to teach you how to clean the same data set using OpenRefine in another video.

But the tool that I'm gonna be using for this one is called Dataprep. DataPrep by Trifacta is a tool that is part of the Google Cloud suite of tools to help you analyze process and clean data. Dataprep is really tailored to massive data sets, but it's also a very interesting application to explore and to do like quick visualization and quick cleaning of data sets. But it's really powerful when you're using Dataprep for really, really massive big data sets for like millions and millions of rows and you want to do like perform cleaning routines that are the same. But we can also use for this small data set that Sarah put together.

This is a small data set is that it has only six thousand, about six thousand almost seven thousand rows, so it's not that big. But let me show you, so as soon as you enter on cloud.google.com/dataprep you can go to console. It will ask you to create an account and then as soon as you create an account you go to a screen like this where it loads the application and then you can start using Dataprep. So to do that, we're going to import data here and to import data we're going to choose a file. Here's the Excel file, we're gonna put this to download on the forums, don't worry about it. You just click here and then it'll start, it'll start uploading for the tool.

Right? So it automatically recognizes that it's an Excel file. And then it tells me that there is one sheet in the file and I want to load this specific sheet. And I'm just going to click "Import & Wrangle" so they can start working on the file right away. As soon as I do that Dataprep greets me with its interface and it's fairly intuitive. You don't have to worry about a lot of the options that are here for this complexity of data set that we want, but some things are worth noting. The first one is like this visual representation of the columns here at the top. Dataprep will always tell you if there are missing columns or mismatch values so you can always find this at a click. So right now it tells me that column 2 here has over four thousand missing values here. So those are all these blank rows and it tells me right away here that those are missing. It also tells me on column 4 there are mismatched values. Those are values that are not of the same data type. It helps me find those values.

So what I want to do here first is just remember that this is what I want. I want a column with the plans, I want another column with the counties, and then enrollment, month and year. Doesn't have to be in the same order but this is what we want, right? So getting back here to Dataprep. What I'm gonna do first is to extract that. You're right. So I want a column with the year. So I want to take this year out and create a column out of this. So, I'm just going to select here 2009 and Dataprep will automatically here on the right suggest a few actions that I might want to do. So what I want here is to extract the values, right? So I want to extract values that are matching digit four, so four digits here.

Not like explicitly or literally 2009, I want four digits because then I get all of the years that are here in this data set. So this is correct, this is the thing that I wanted too. And see that in blue it highlights my original column, in yellow it previews the column that's going to create. And it's going to create a column that has values 2008, '09, '10, '11, '12, '13 and also '14. It even gives me like how many values so, twelve 2018, that's point eighteen percent of this column and twelve 2009, so on and so forth. So I want to, I'm going to go ahead and click on "add".

[00:07:30] So it added a column here. I'm not going to worry about the names, I'm going to rename all of the columns later. But now I want to extract the month, right? And to extract a month I, I'm going to I'm going to go and just select a month here. But it's not gonna be that easy, right? So with the year it was fairly easy but with a month here what we're going to do is really to select the extract here but we're going to edit what's going on, and you can select any of the actions.

So the column to extract from it is really column 2. And we're going to do, like a custom text pattern. So I'm going to delete this and then I'm gonna say that's gonna start extracting from "NYS", right? So there's always a NYS here before the month, and then it's gonna be always before these four digits. Remember when we extracted the four digits so we're gonna type here just the way they're indicating here. So digit and then four. All right. So here it shows that it's extracting everything that I need actually more than I need because it's extracting as well the spaces here that are around it and then the comma. But we're gonna deal with that soon.

So I'm gonna add, and then it creates another column with the months. And I'm going to do with this empty space here, I'm just going to select this empty space. And see how Dataprep already suggests that we replace the values here, even this space before January and after January with nothing. So we're just going to delete those empty spaces for me. So I'm gonna select this one, now I'm going to click add. And then you see that removed all of that. Now I'm going to remove the comma too, just find a comma and replace with nothing, so you're gonna delete the comma. All right, so we have like a column for months and a column for years.

But now see that we have the values here. So January is here and then we scroll down January is here again with a different year. Then we'll keep scrolling down, we see

January again. So we want to, we want to make sure that all of these values here are filled with January and all these values here are filled with 2009, until they get to 2010 and then they get filled down again until they find the other year 2011. So on and so forth. This is called a Fill Down and we can do this automatically in Dataprep just by using a new step.

So we're going to click on "new step" and then we're going to type here "window". And then there is a former called "fill", and then we open parenthesis and then we select a column. So we're gonna do first in the column 5, which is the column at the moment that we're using for months. So I'm just going to type "fill" and then "column5". And then I'm going to type here source number which is the row number in the original source file here. So you type this and see that the blue column here which is the source becomes this yellow column here. So January gets filled down until it finds a different value which is February. So if I keep scrolling down until we find a different month it finds February then until the end of the data set.

So I'm just going to add this here and then it will create another column here for us. This will be our month column and I'm going to do the same with the year column so "new step", "window", and then I'm gonna put "fill", and that's the column 1. Here I don't need that, "column1" and then source number here, "sourcerownumber". And then, yes it creates here another column for me that will be my year column. Now click "add". All right, I have the county here, the enrollment here, the month, and the year, and the plan.

So it's starting to look like this thing that we want, but we still need to do a few things. So the first thing that we're going to do after we created those columns we might want now to rename all of them. So I'm going to rename this one, health plan. So, again I just clicked on here on this arrow pointing downwards. Click "rename", and then I'm going to rename just to "plan". I might want to do another thing with this one too, but I'm just gonna rename that to plan. Same thing. This one actually I'm going to remove because I don't need it. I'm also going to remove this one, "delete". Rename this to "county". This one rename to "enrollments". All right. And this one to "month", and this one to "year".

All right, so now look at the visual representation here at the top. We can find the missing values too and all of the mismatched values. But we're going to remove, start removing like things like the total, right? We don't need this total, so we're just going to highlight it and ask Dataprep to delete all of the rows that are matching "total" because we don't need that.

Now we might want to do the same here with the health care plans, with the Medicaid plans to fill down, right? Because we want this to be about the same plan that is here, because if you take a look at the original table the group that Medicaid plans together and have empty spaces to indicate they are the same. So we need to make sure that those are filled down with the names of the plans too. So we're gonna go ahead and do that. And if you remember, this is a new step. "Window", the formula is "fill", and then we go to "plan" which is the name of the column now, and then "sourcerownumber". And then it creates a new column here that has all of the values filled down.

I'm wanting, I want to delete this too and then move this one to the beginning. I'm going to add and rename this one to "plan" again. All right, now I want to remove like all of the missing values here because they, I mean they either have totals or they're about rows that are not part of my data set, my final data sets. I'm just going to remove them. I'm also going to take a look at mismatched values here, those mismatched values are those that are not numbers. So I'm going to remove the mismatched values too, and there are missing values here too that I want to remove.

So this is this is looking much more like the data set that we have here, and we completed this in only twenty one steps. So this is, this is the final clean data set. So now what you should do is, if you want to export this to a CSV file you actually select this option here to "run job". So Dataprep will use its cloud infrastructure to get this recipe that you just created and then apply it to the data sets that you imported. That's why it's so much more powerful when you're dealing with massive data sets, but it can also do this with the small one. You click on "run job", and then you check if everything is ok.

There are options here that you might want to select, for example I want to select that it will create a CSV file and I have more options here that "include headers as first row on creation" and then I update this. I'm going to select a location where it's going to save on my cloud account. I select the regions and then I run the job. And as soon as the job is done you will get a file a CSV file that is like this, I have it here with me as well. So here is a CSV file that you might want to load. So this is it for the Dataprep. Go ahead and go to cloud.google.dataprep and try it out.