Dataprep: Qwik Start

GSP105

Overview

Cloud Dataprep by Trifacta is an intelligent data service for visually exploring, cleaning, and preparing data for analysis. Cloud Dataprep is serverless and works at any scale. There is no infrastructure to deploy or manage. Easy data preparation with clicks and no code!
In this lab you use Dataprep to manipulate a dataset. You import datasets, correct mismatched data, transform data, and join data. If this is new to you, you'll know what it all is by the end of this lab.

Setup and Requirements

Before you click the Start Lab button

Read these instructions. Labs are timed and you cannot pause them. The timer, which starts when you click **Start Lab**, shows how long Google Cloud resources will be made available to you.

This Qwiklabs hands-on lab lets you do the lab activities yourself in a real cloud environment, not in a simulation or demo environment. It does so by giving you new, temporary credentials that you use to sign in and access Google Cloud for the duration of the lab.

What you need

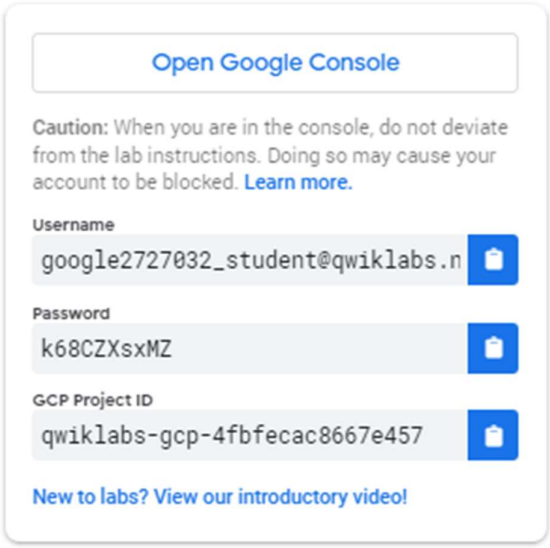
To complete this lab, you need:

- Access to a standard internet browser (Chrome browser recommended).
- Time to complete the lab.
  Note: If you already have your own personal Google Cloud account or project, do not use it for this lab.

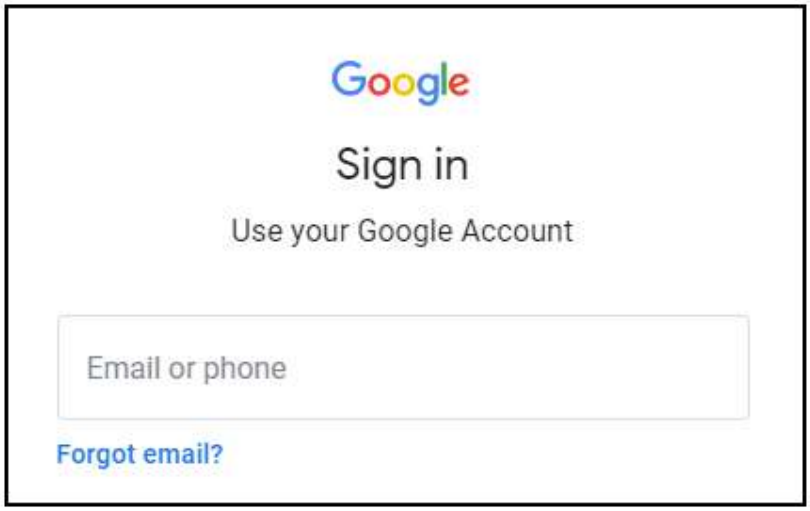
Note: If you are using a Pixelbook, open an Incognito window to run this lab.

How to start your lab and sign in to the Google Cloud Console

1. Click the **Start Lab** button. If you need to pay for the lab, a pop-up opens for you to select your payment method. On the left is a panel populated with the temporary credentials that you must use for this lab.

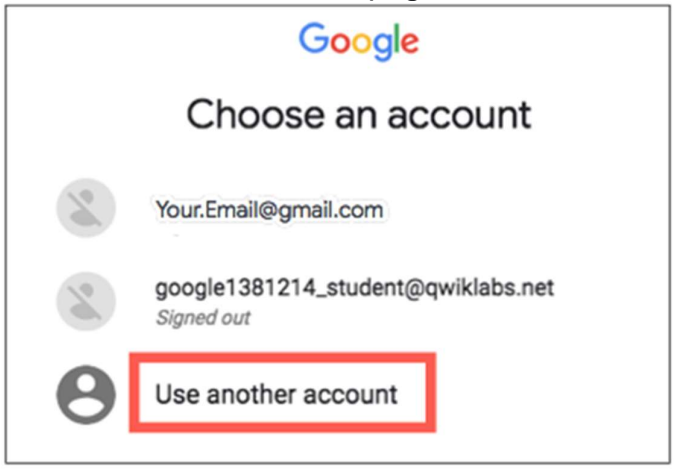


2. Copy the username, and then click **Open Google Console**. The lab spins up resources, and then opens another tab that shows the **Sign in** page.



Tip: Open the tabs in separate windows, side-by-side.

If you see the **Choose an account** page, click **Use Another Account**.



3. In the **Sign in** page, paste the username that you copied from the Connection Details panel. Then copy and paste the password.

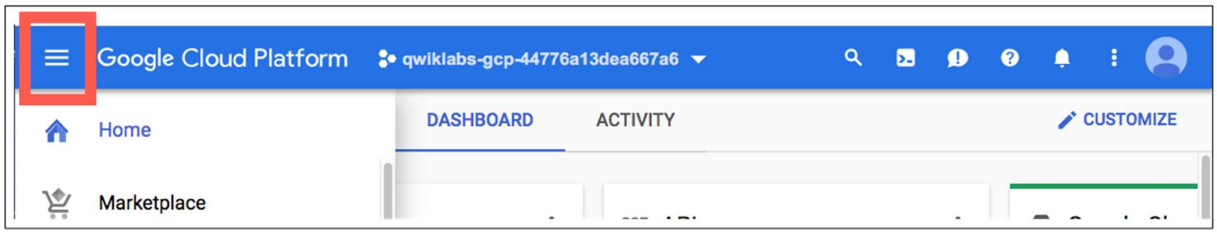
   Important: You must use the credentials from the Connection Details panel. Do not use your Qwiklabs credentials. If you have your own Google Cloud account, do not use it for this lab (avoids incurring charges).

4. Click through the subsequent pages:

   - Accept the terms and conditions.
   - Do not add recovery options or two-factor authentication (because this is a temporary account).
   - Do not sign up for free trials.

After a few moments, the Cloud Console opens in this tab.

Note: You can view the menu with a list of Google Cloud Products and Services by clicking the **Navigation menu** at the top-left.

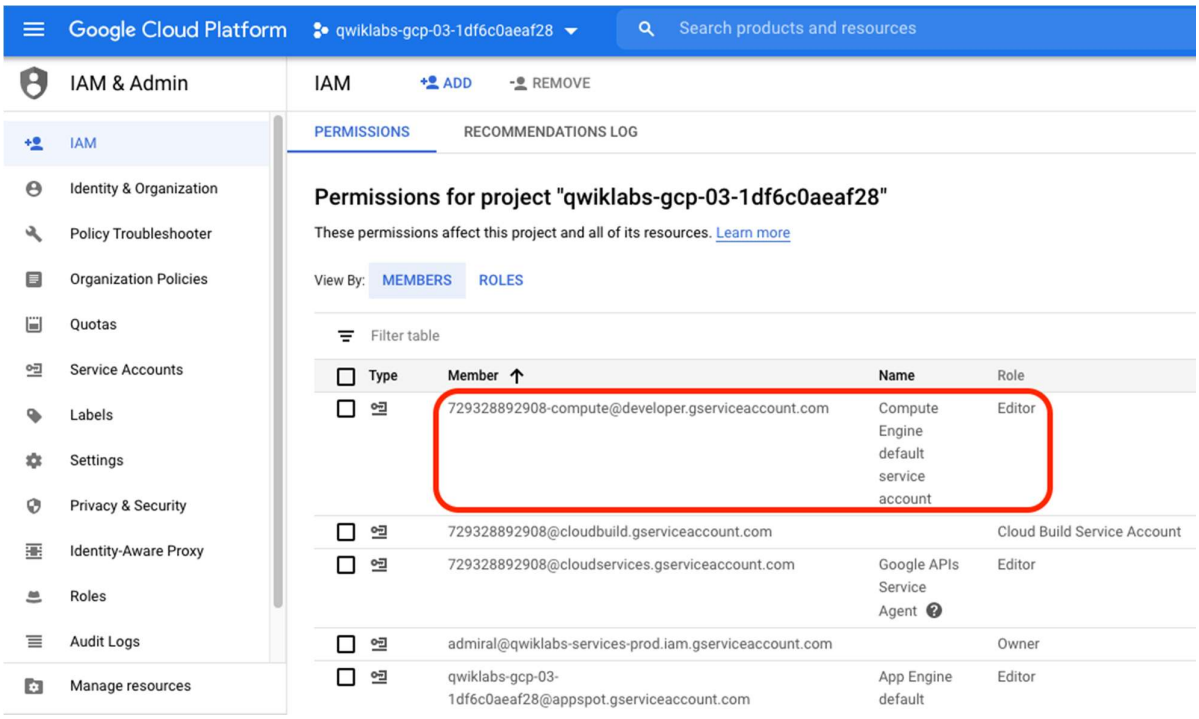


Check project permissions

Before you begin your work on Google Cloud, you need to ensure that your project has the correct permissions within Identity and Access Management (IAM).

1. In the Google Cloud console, on the **Navigation menu** (), click **IAM & Admin** > **IAM**.

2. Confirm that the default compute Service Account `{project-number}-compute@developer.gserviceaccount.com` is present and has the `editor` role assigned. The account prefix is the project number, which you can find on **Navigation menu** > **Home**.



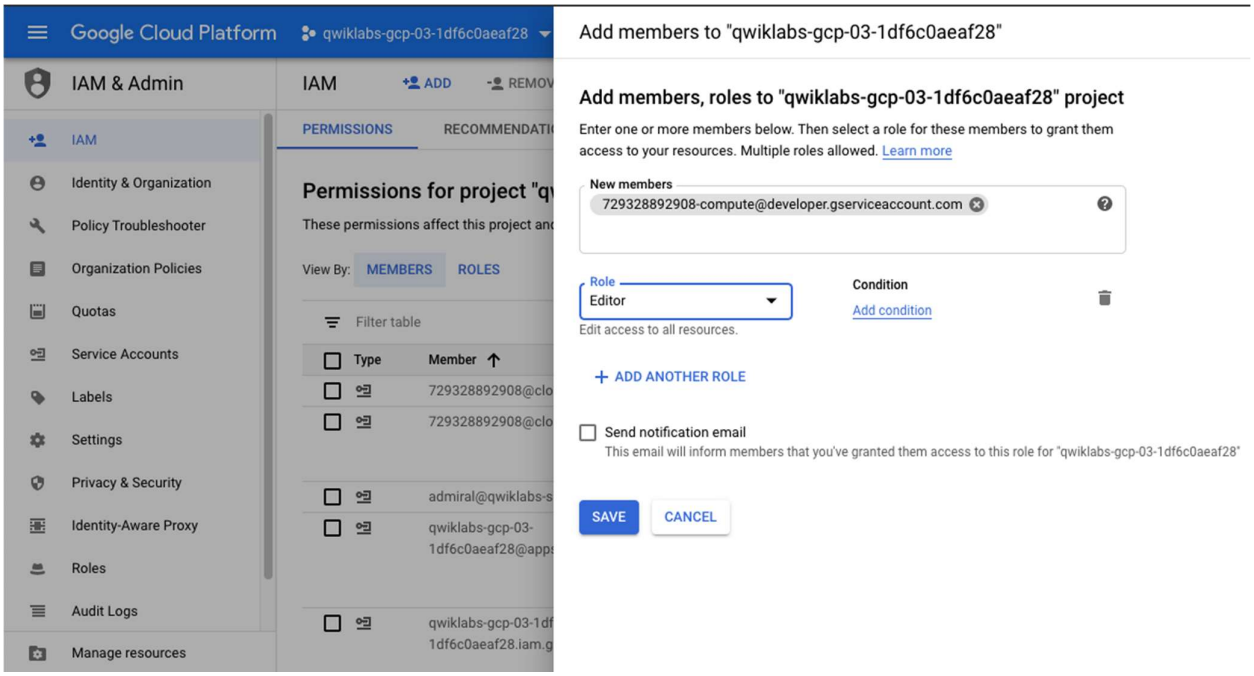
If the account is not present in IAM or does not have the `editor` role, follow the steps below to assign the required role.

- In the Google Cloud console, on the **Navigation menu**, click **Home**.
- Copy the project number (e.g. `729328892908`).
- On the **Navigation menu**, click **IAM & Admin** > **IAM**.
- At the top of the **IAM** page, click **Add**.
- For **New members**, type:

```
{project-number}-compute@developer.gserviceaccount.com
```

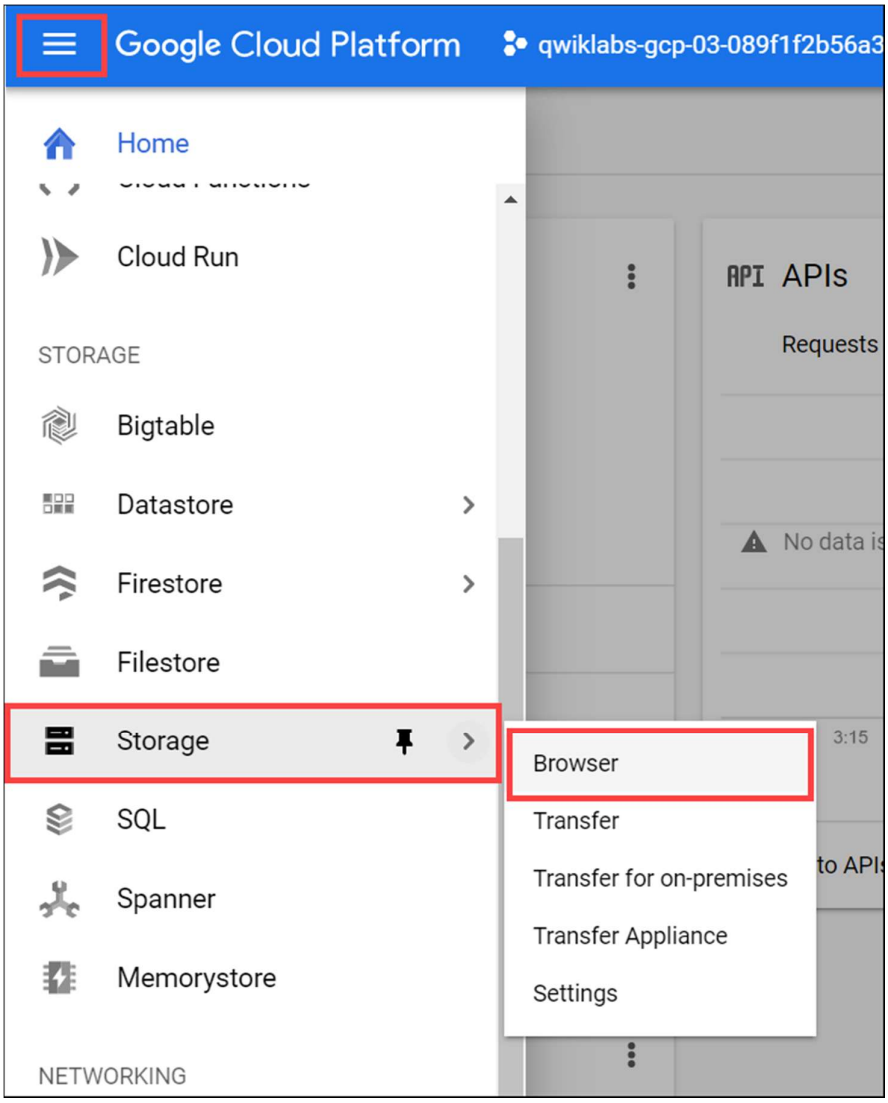
Replace `{project-number}` with your project number.

- For **Role**, select **Project** (or Basic) > **Editor**. Click **Save**.


Google Cloud Platform
qwiklabs-gcp-03-1df6c0aeaf28
IAM & Admin
IAM
ADD
REMOV
PERMISSIONS
RECOMMENDATI
Permissions for project "q
These permissions affect this project an
View By:
MEMBERS
ROLES
Filter table
Type
Member
729328892908@clo
729328892908@clo
admiral@qwiklabs-s
qwiklabs-gcp-03-
1df6c0aeaf28@apps
qwiklabs-gcp-03-1df
1df6c0aeaf28.iam.g
IAM
Identity & Organization
Policy Troubleshooter
Organization Policies
Quotas
Service Accounts
Labels
Settings
Privacy & Security
Identity-Aware Proxy
Roles
Audit Logs
Manage resources
Add members to "qwiklabs-gcp-03-1df6c0aeaf28"
Add members, roles to "qwiklabs-gcp-03-1df6c0aeaf28" project
Enter one or more members below. Then select a role for these members to grant them access to your resources. Multiple roles allowed. Learn more
New members
729328892908-compute@developer.gserviceaccount.com
Role
Editor
Edit access to all resources.
Condition
Add condition
ADD ANOTHER ROLE
Send notification email
This email will inform members that you've granted them access to this role for "qwiklabs-gcp-03-1df6c0aeaf28"
SAVE
CANCEL

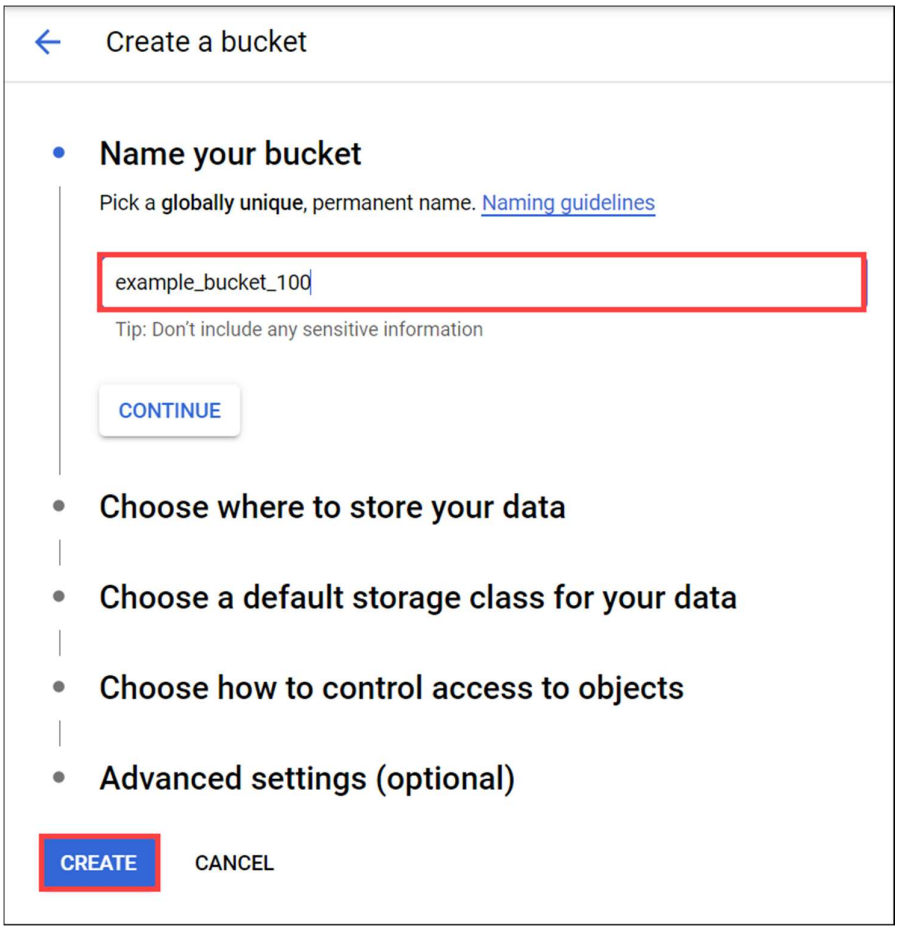
Create a Cloud Storage bucket in your project

1. In the Cloud Platform Console, select **Navigation menu** > **Storage** > **Browser**.



2. Click **Create bucket**.

3. In the **Create a bucket** dialog, **Name** the bucket a unique name. Skip the other steps to leave those settings at their default value.



For more information about naming buckets, see bucketname requirements.

4. Click **Create**.

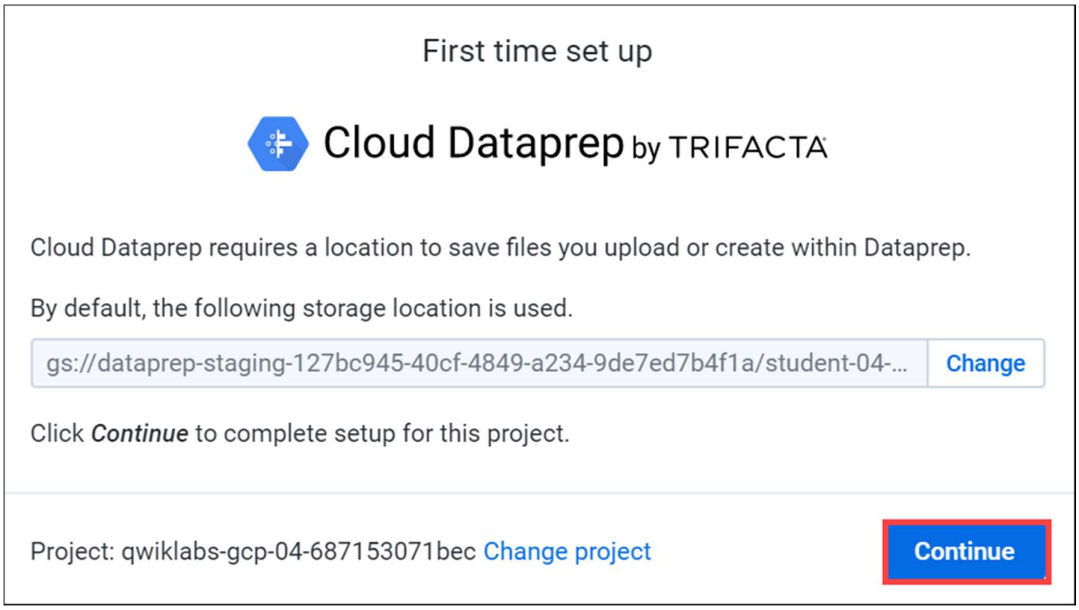
You created your bucket. Remember the bucket name for later steps.

Test Completed Task

Click **Check my progress** to verify your performed task. If you have successfully created Cloud Storage bucket, you see an assessment score.

Initialize Cloud Dataprep

1. Select **Navigation menu** > **Dataprep**.
2. Check to accept the Google Dataprep Terms of Service, then click **ACCEPT**.
3. Check to authorize sharing your account information with Trifacta, then click **Agree and Continue**.
4. Click **Allow** to allow Trifacta to access project data.
5. Click your Google Cloud username to sign in to Cloud Dataprep by Trifacta. Your Google Cloud username is **Username** in the left panel in your lab.
6. Click **Allow** to grant Cloud Dataprep access to your Google Cloud lab account.
7. Check to agree to Trifacta Terms of Service, and then click **Accept**.
8. Click **Continue** on the "First time set up" screen to create the default storage location.



Dataprep opens in a new browser tab. A tutorial will launch, asking you to select datasets. Quit out of this screen by clicking **Cancel** or exiting out.

Click on the Dataprep icon on the top left corner to go to the home screen.

Test Completed Task

Click **Check my progress** to verify your performed task. If you have successfully initialized Cloud Dataprep with default storage location, you see an assessment score.

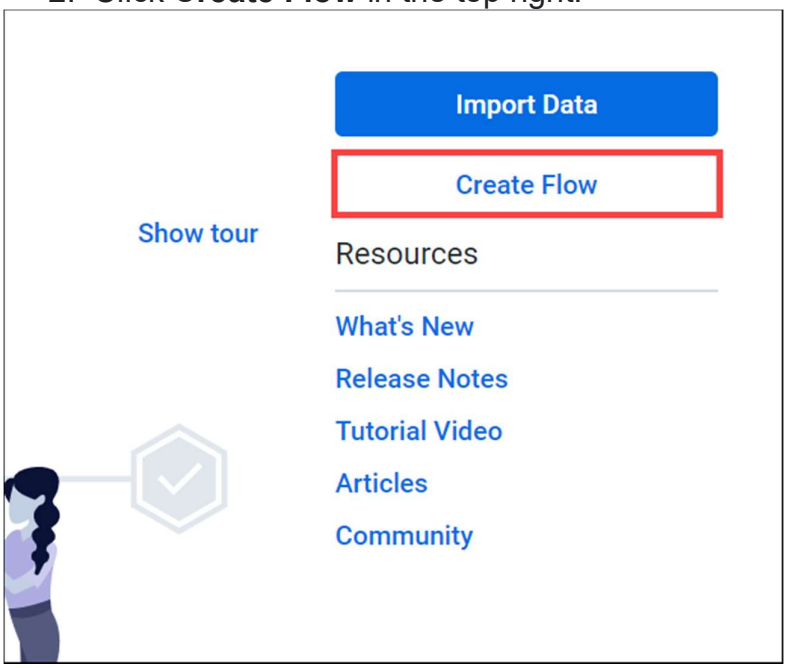
Initialize Cloud Dataprep

Check my progress

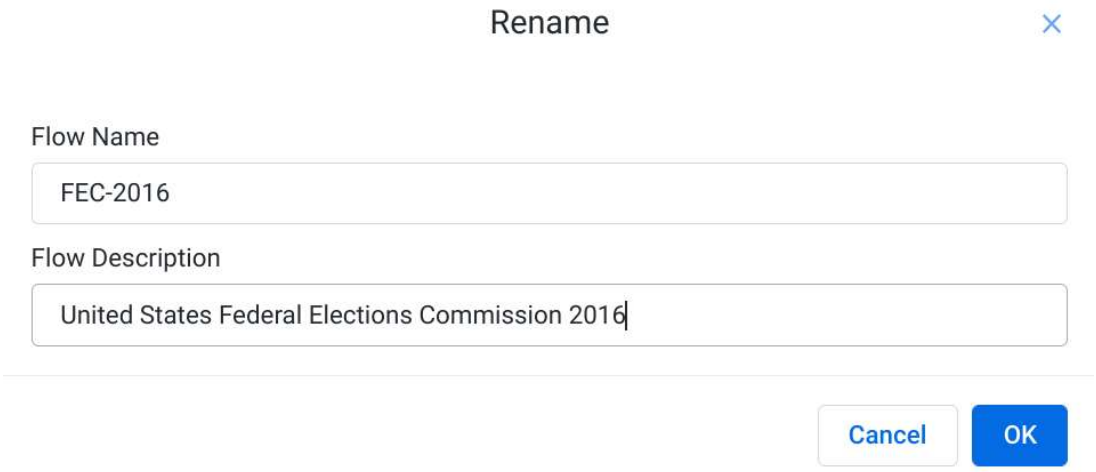
Create a flow

Cloud Dataprep uses a `flow` workspace to access and manipulate datasets.

1. Click on flows in left navigation rail.
2. Click **Create Flow** in the top right:



2. Choose Blank.
3. click on untitled to change flow name and description.
4. Name and describe the flow. Since this lab uses 2016 data from the [United States Federal Elections Commission 2016], name the flow "FEC-2016", and the describe the flow as "United States Federal Elections Commission 2016".



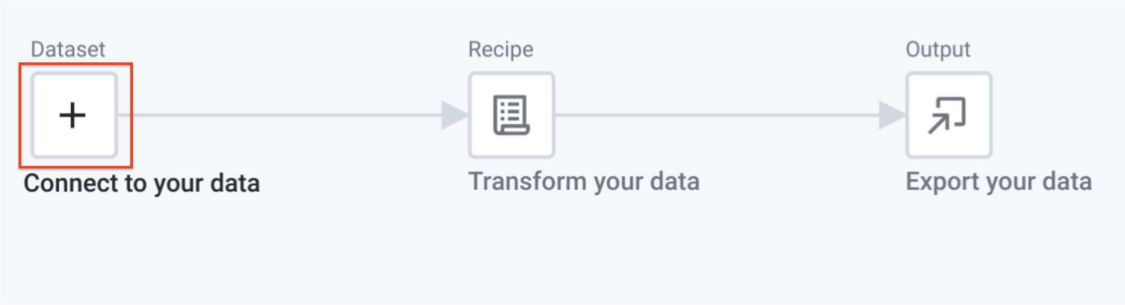
3. Click **OK**.

The FEC-2016 flow page opens.

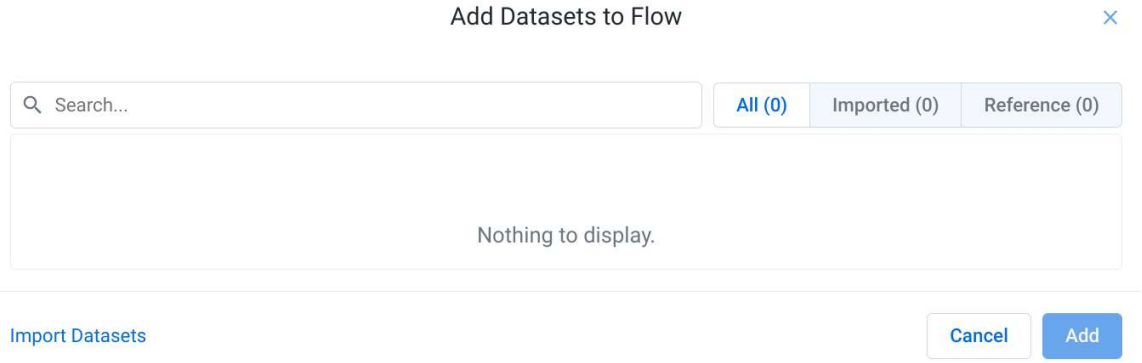
Import datasets

In this section you import and add data to the FEC-2016 flow.

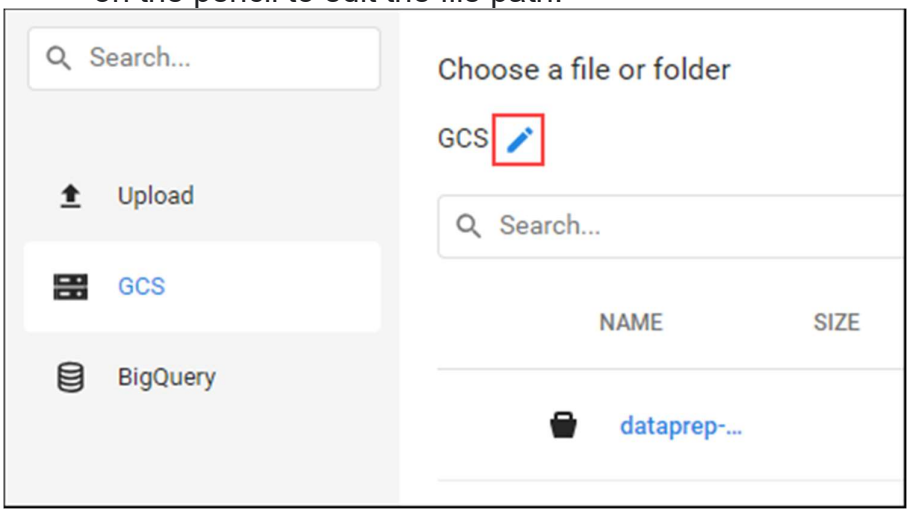
1. Click the **Add Icon** in the Dataset box.



2. In the **Add Datasets to Flow** dialog box, select **Import Datasets**.



3. In the left menu pane, select **GCS** to import datasets from Cloud Storage, then click on the pencil to edit the file path.



4. Type `gs://spls/gsp105` in the **Choose a file or folder** text box, then click **Go**.

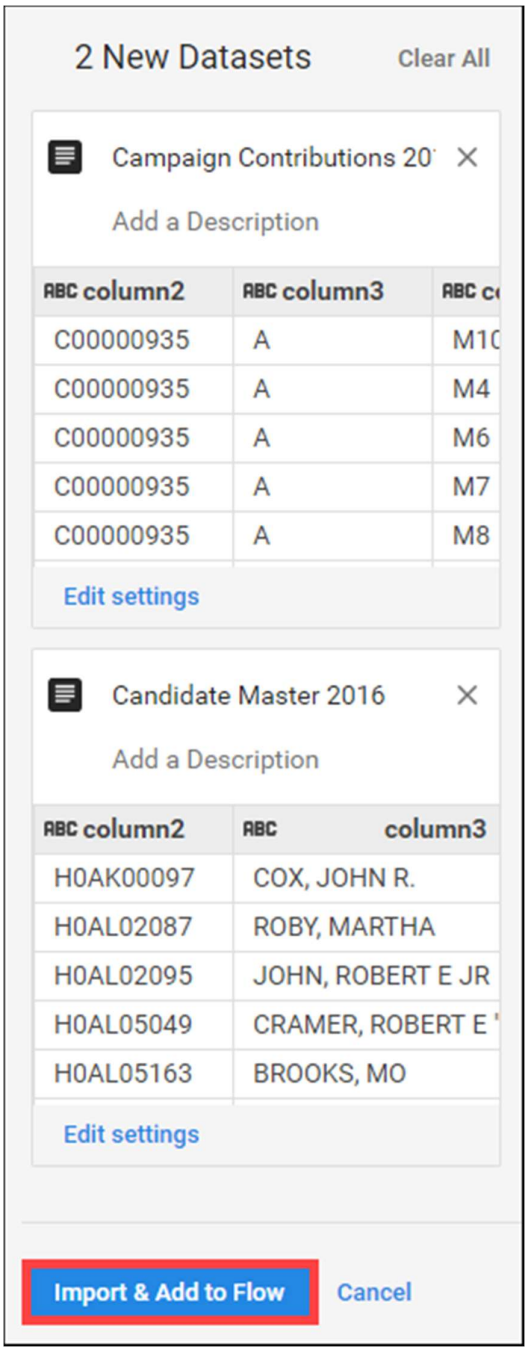
You may have to widen the browser window to see the **Go** and **Cancel** buttons.

5. Click **us-fec/**.

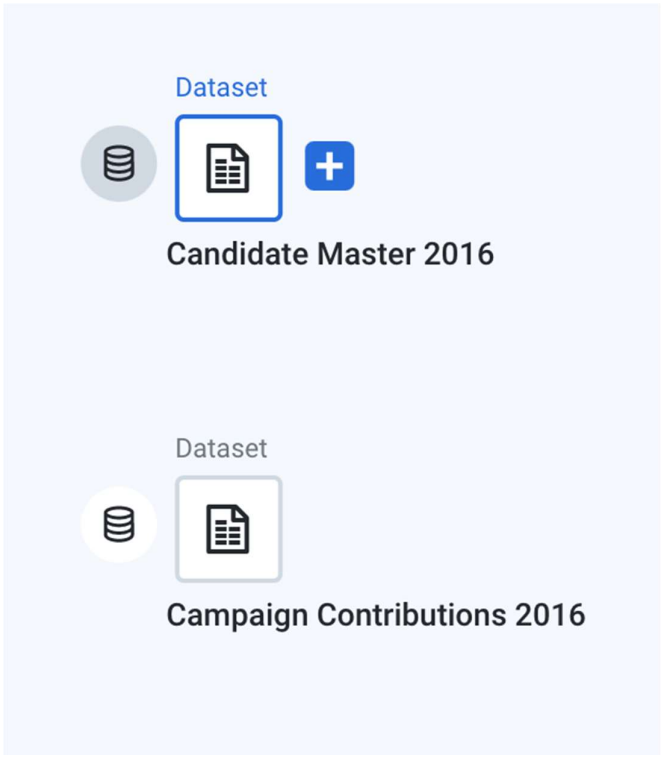
6. Click the **+** icon next to `cn-2016.txt` to create a dataset shown in the right pane. Click on the title in the dataset and rename it "Candidate Master 2016".

7. In the same way add the `itcont-2016.txt` dataset, and rename it "Campaign Contributions 2016".

8. After both datasets are listed in the right pane, click **Import & Add to Flow**.

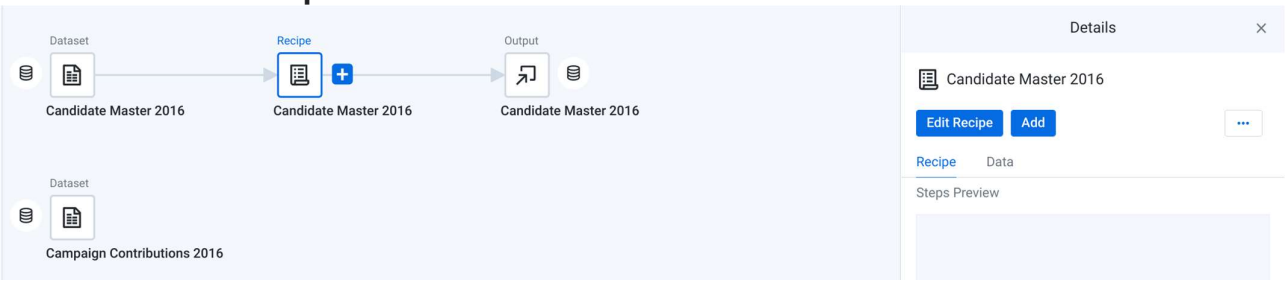


You see both datasets listed as a flow.

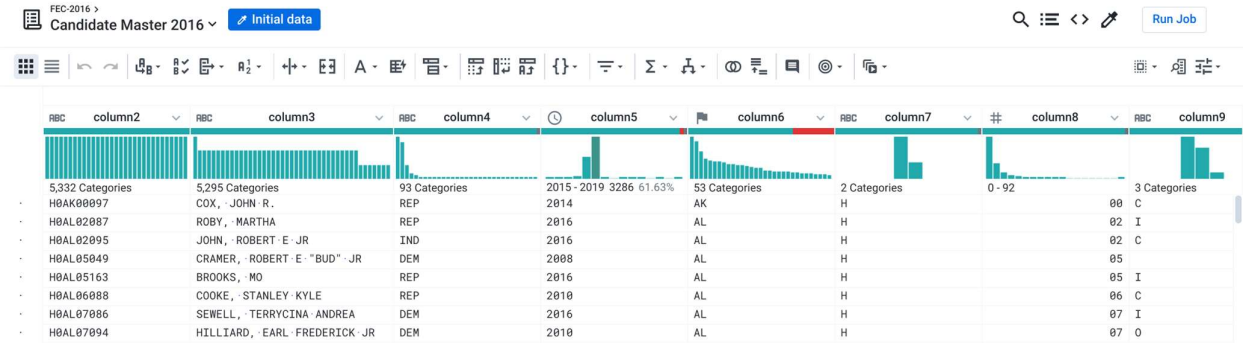

Dataset
Candidate Master 2016
Dataset
Campaign Contributions 2016

Prep the candidate file

1. By default, the Candidate Master 2016 dataset is selected. In the right pane, click **Edit Recipe**.

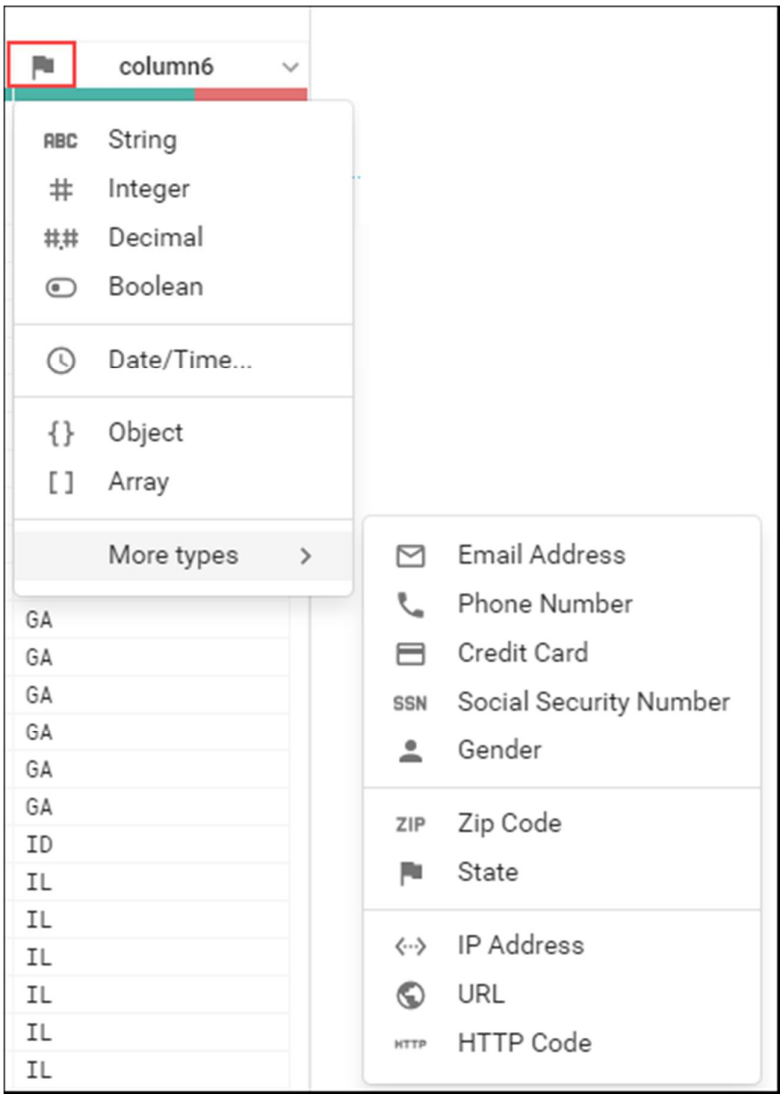


The Candidate Master 2016 Transformer page opens in the grid view.

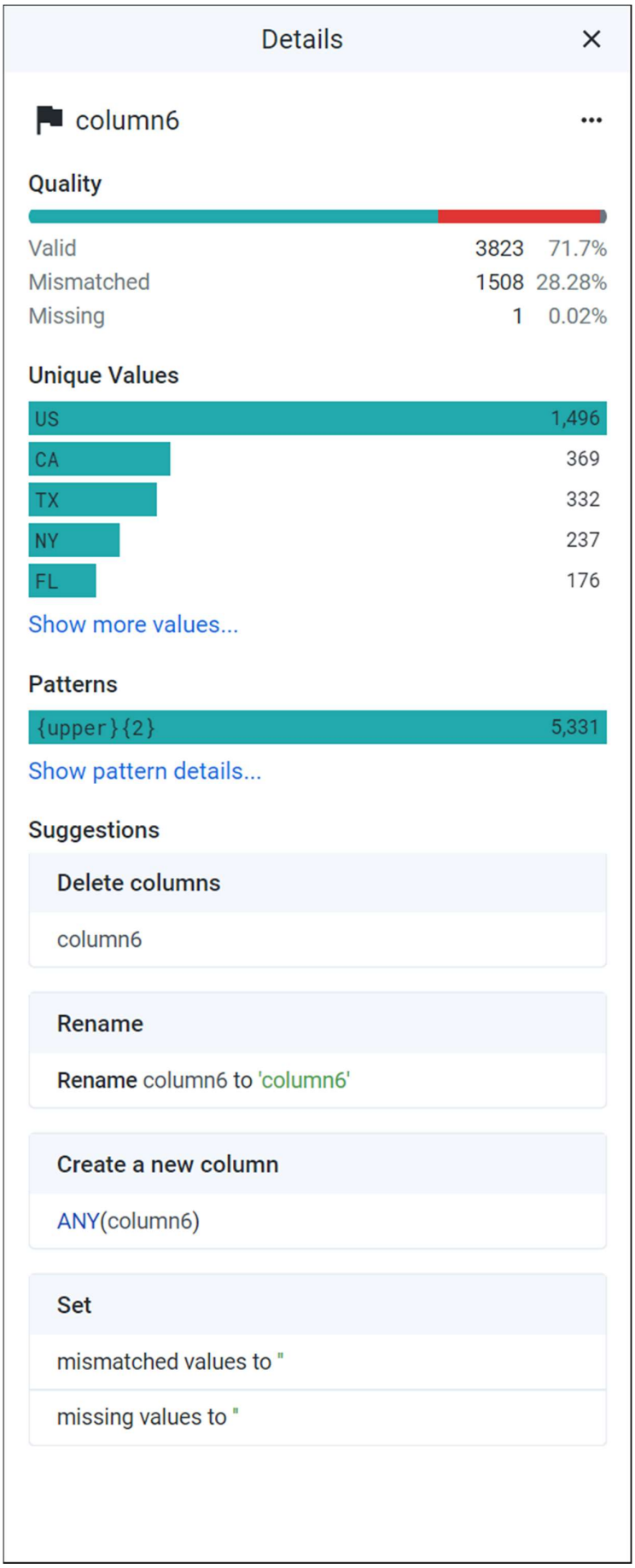


The Transformer page is where you build your transformation recipe and see the results applied to the sample. When you are satisfied with what you see, execute the job against your dataset.

Each of the column heads have a Name and value that specify the data type. Data types are shown when you click the flag icon:



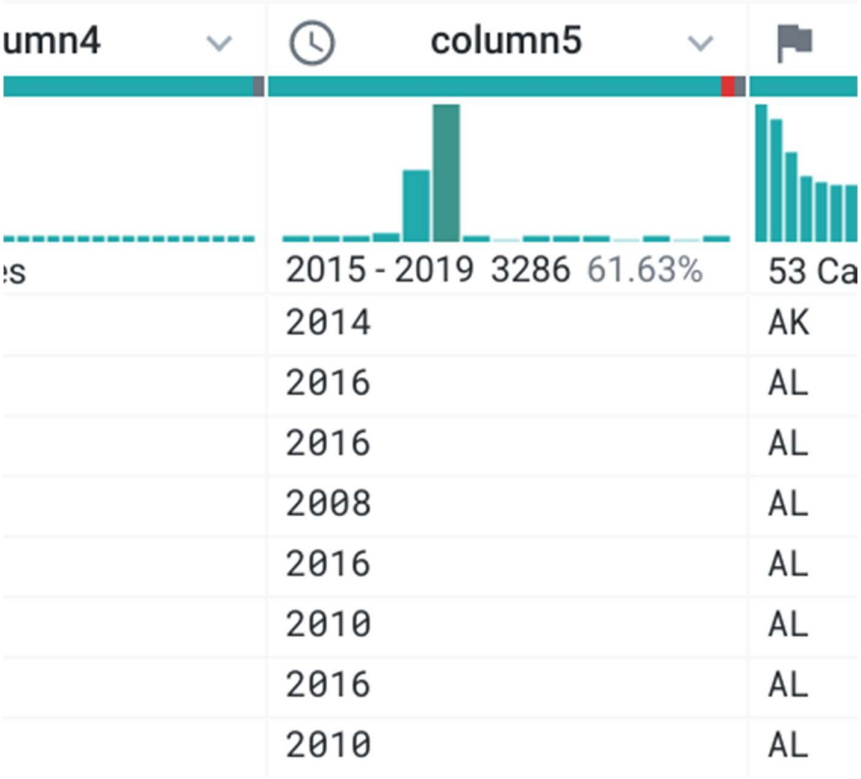
Also, when you click the column name, a **Details** panel opens on the right:



Click **X** in the top right of the Details panel to close the Details panel.

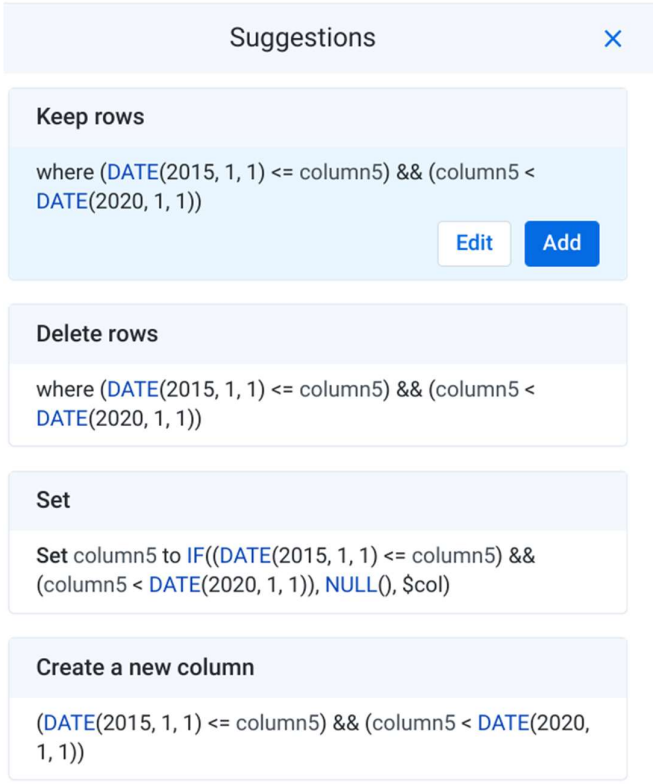
In the following steps, you explore data in the grid view and apply transformation steps to your recipe.

2. Column5 provides data from 1990-2064. Widen column5 (like you would on a spreadsheet) to separate each year. Click to select the tallest bin, which represents year 2015.



This creates a step where these values are selected.

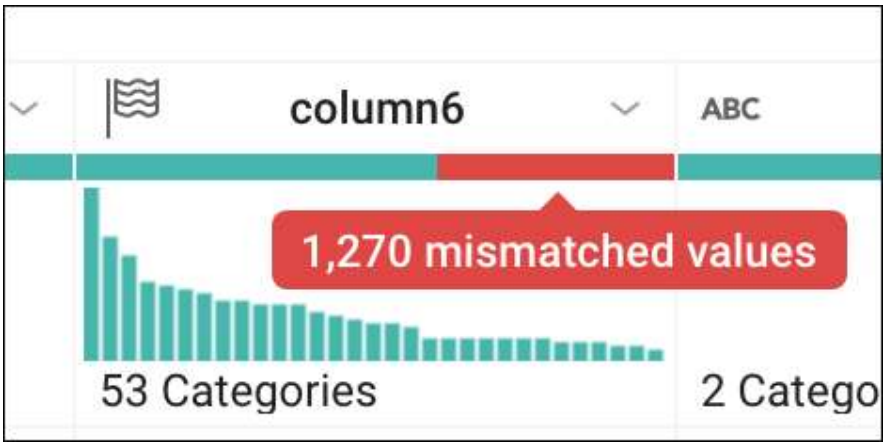
3. Click on a dot at start of any row. In the **Suggestions** panel on the right, in the **Keep rows** section, click **Add** to add this step your recipe.



The Recipe panel on the right now has the following step:

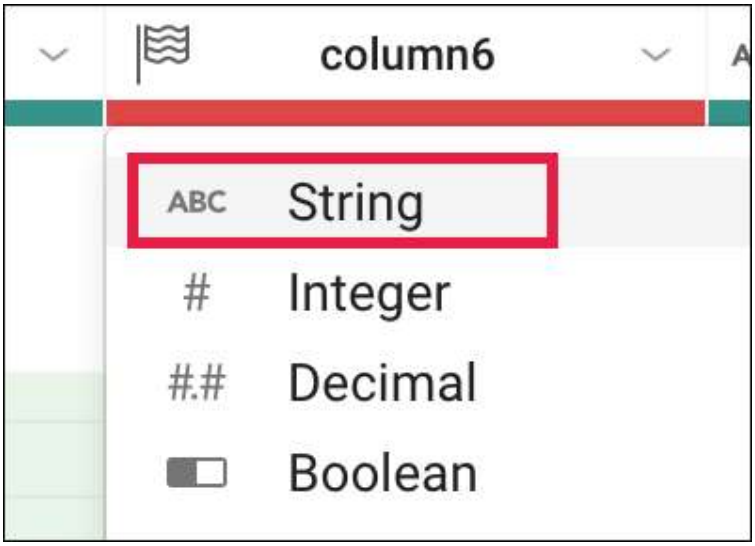
```
Keep rows where(DATE(2015, 1, 1) <= column5) && (column5 < DATE(2020, 1, 1))
```

4. In Column6 (State), hover over and click on the mismatched (red) portion of the header to select the mismatched rows.



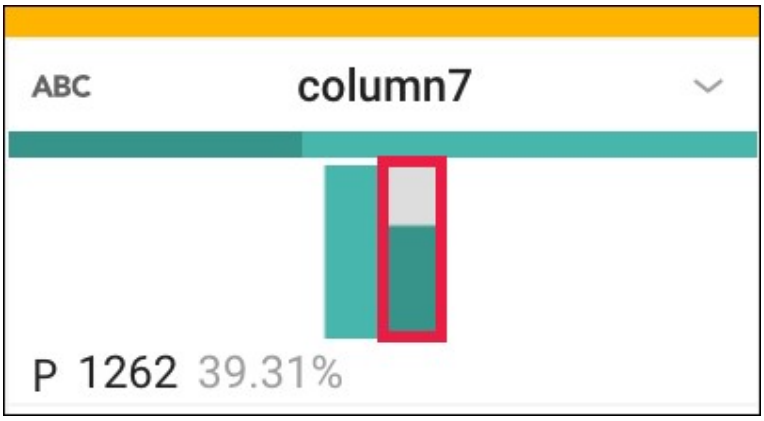
Scroll down to find the mismatched values and notice how most of these records have the value "P" in column7, and "US" in column6. The mismatch occurs because column6 is marked as a "State" column (indicated by the flag icon), but there are non-state (such as "US") values.

5. To correct the mismatch, click **X** in the top of the Suggestions panel to cancel the transformation, then click on the flag icon in Column6 and change it to a "String" column.

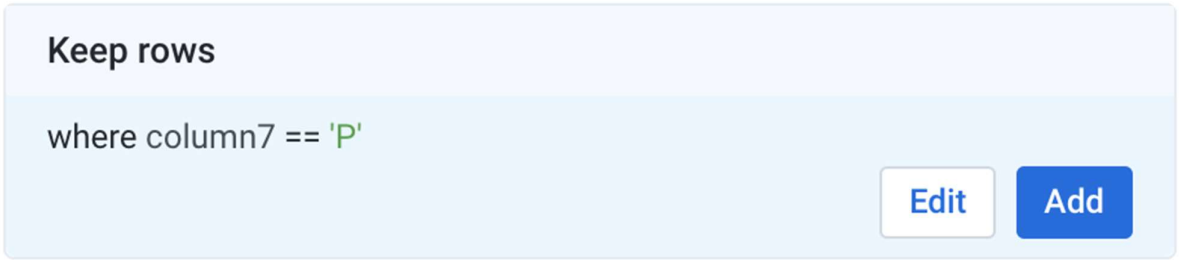


There is no longer a mismatch and the column marker is now green.

6. Filter on just the presidential candidates, which are those records that have the value "P" in column7. In the histogram for column7, hover over the two bins to see which is "H" and which is "P". Click the "P" bin.



7. In the right Suggestions panel, click **Add** to accept the step to the recipe.

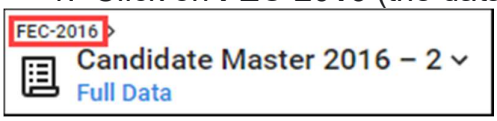


Join the Contributions file

On the Join page, you can add your current dataset to another dataset or recipe based on information that is common to both datasets.

Before you join the Contributions file to the Candidates file, clean up the Contributions file.

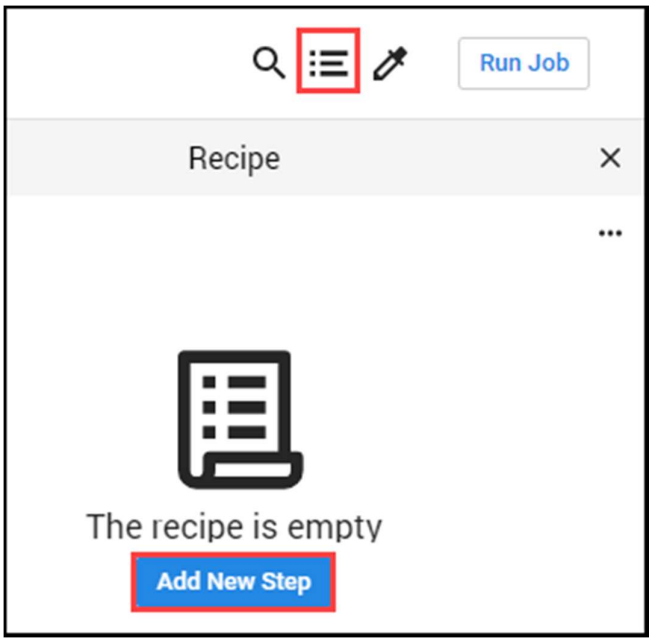
1. Click on **FEC-2016** (the dataset selector) at the top of the grid view page.



2. Click to select the grayed out **Campaign Contributions**.

3. In the right pane, click **Add** > **Recipe**, then click **Edit Recipe**.

4. Click the **recipe** icon at the top right of the page, then click **Add New Step**.

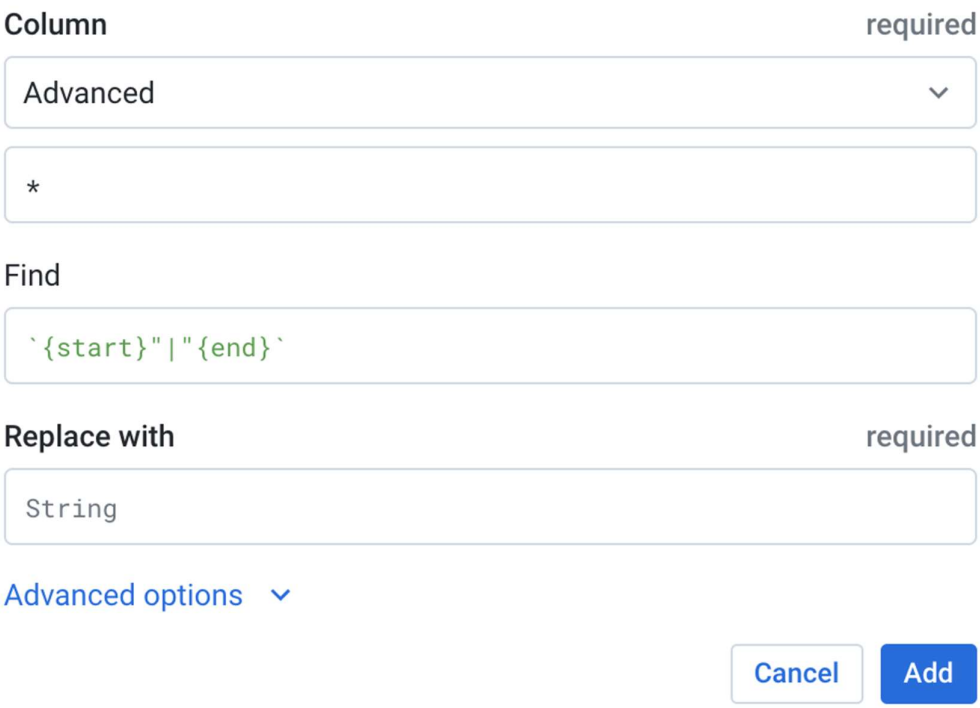


Remove extra delimiters in the dataset.

5. Insert the following Wrangle language command in the Search box:

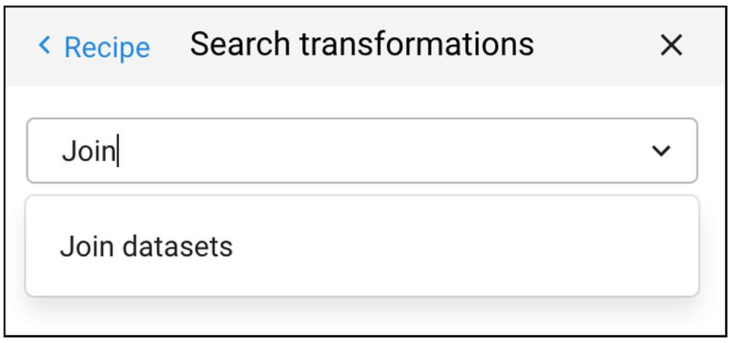
```
replacepatterns col: * with: '' on: `{start}"|"{end}` global: true
```

The Transformation Builder parses the Wrangle command and populates the Find and Replace transformation fields.



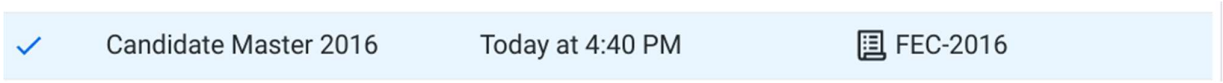
6. Click **Add** to add the transform to the recipe.

7. Add another new step to the recipe. Click **New Step**, then type "Join" in the Search box.

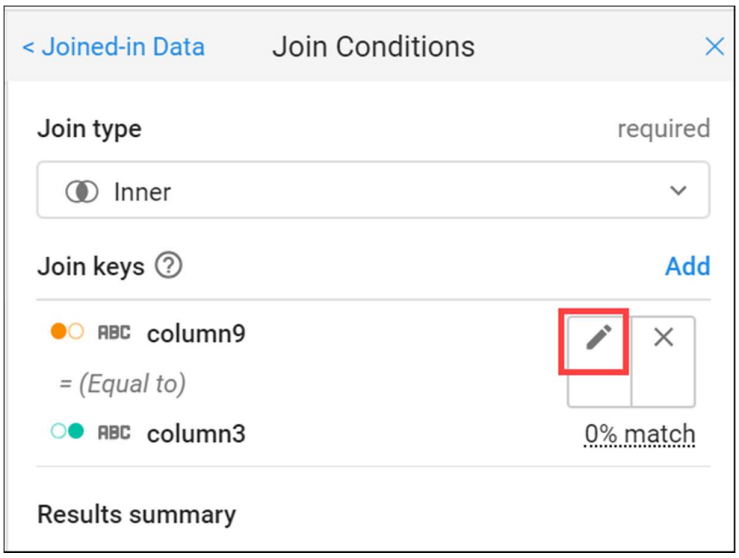


8. Click **Join datasets** to open the Joins page.

9. Click on "Candidate Master 2016" to join with Campaign Contributions 2016-2, then **Accept** in the bottom right.

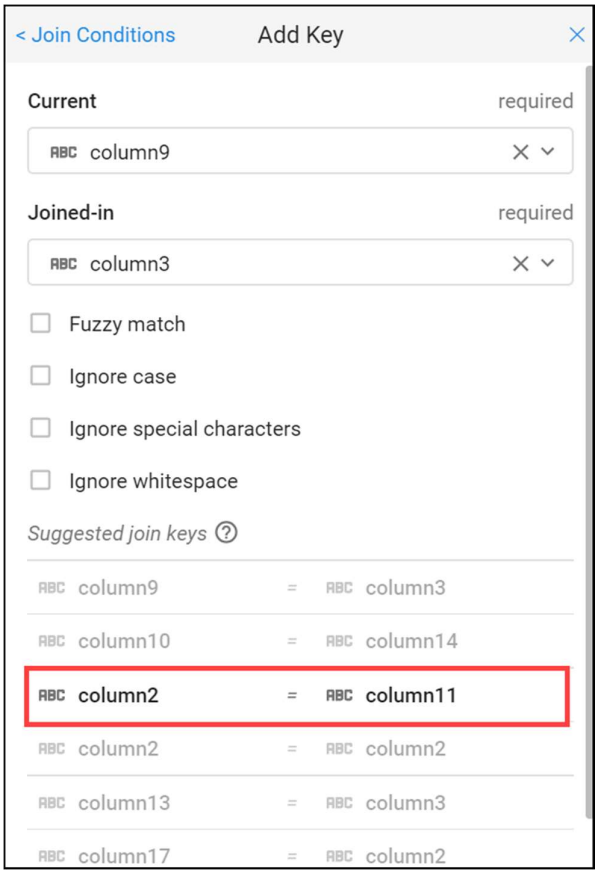


10. Hover in the Join keys section, then click on the pencil (Edit icon).



Dataprep infers common keys. There are many common values that Dataprep suggests as Join Keys.

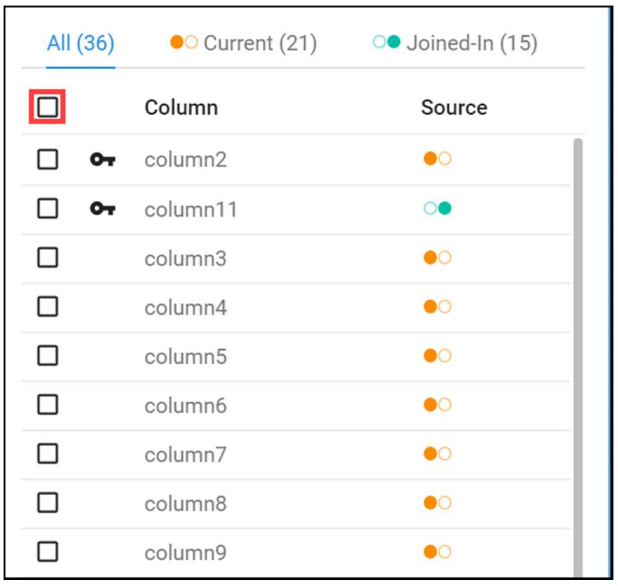
11. In the Add Key panel, in the Suggested join keys section, click **column2 = column11**.



12. Click **Save and Continue**.

Columns 2 and 11 open for your review.

13. Click **Next**, then check the checkbox to the left of the "Columns" label to add all columns of both datasets to the joined dataset.



14. Click **Review**, and then **Add to Recipe** to return to the grid view.

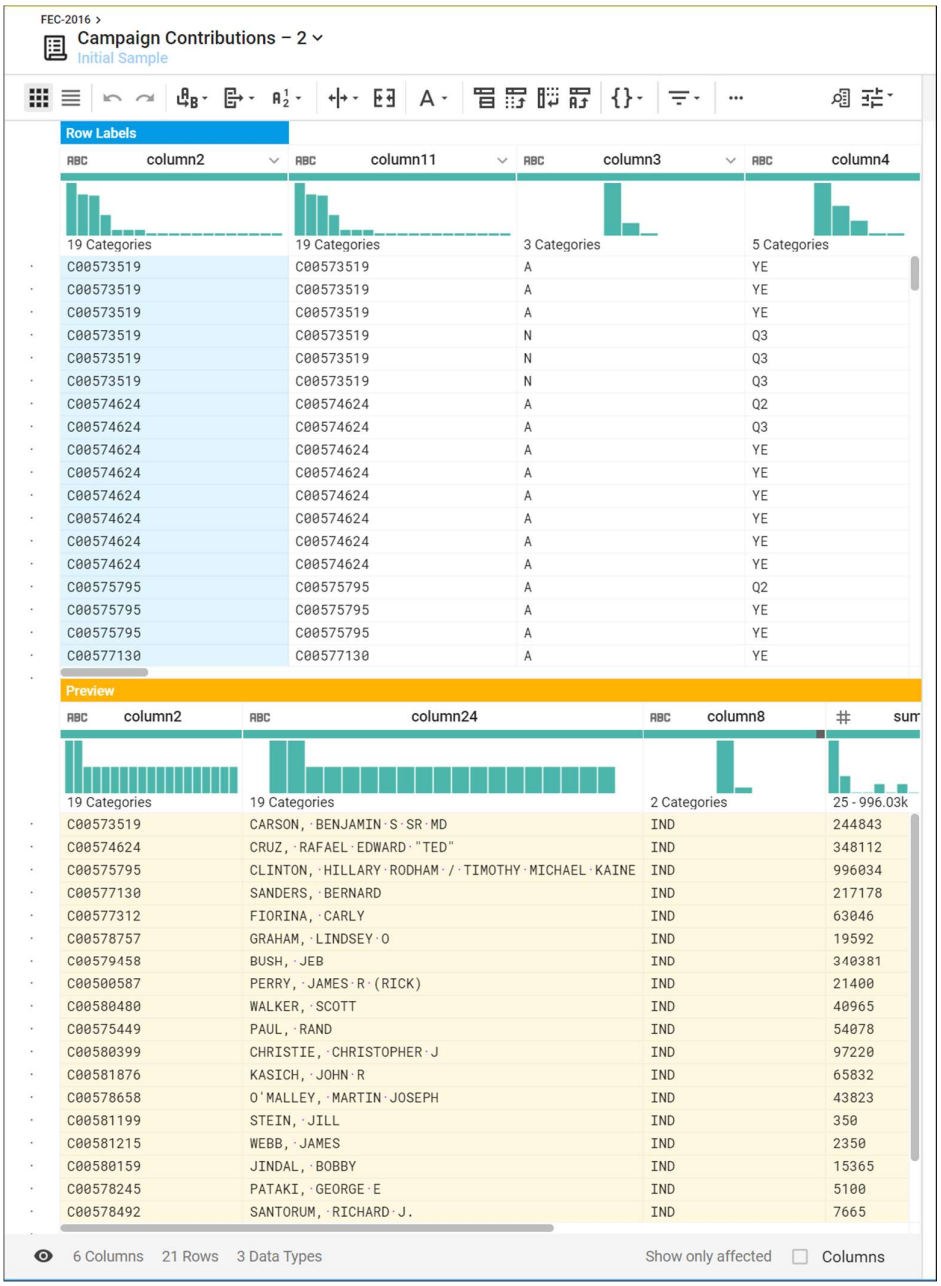
Summary of data

Generate a useful summary by aggregating, averaging, and counting the contributions in Column 16 and grouping the candidates by IDs, names, and party affiliation in Columns 2, 24, 8 respectively.

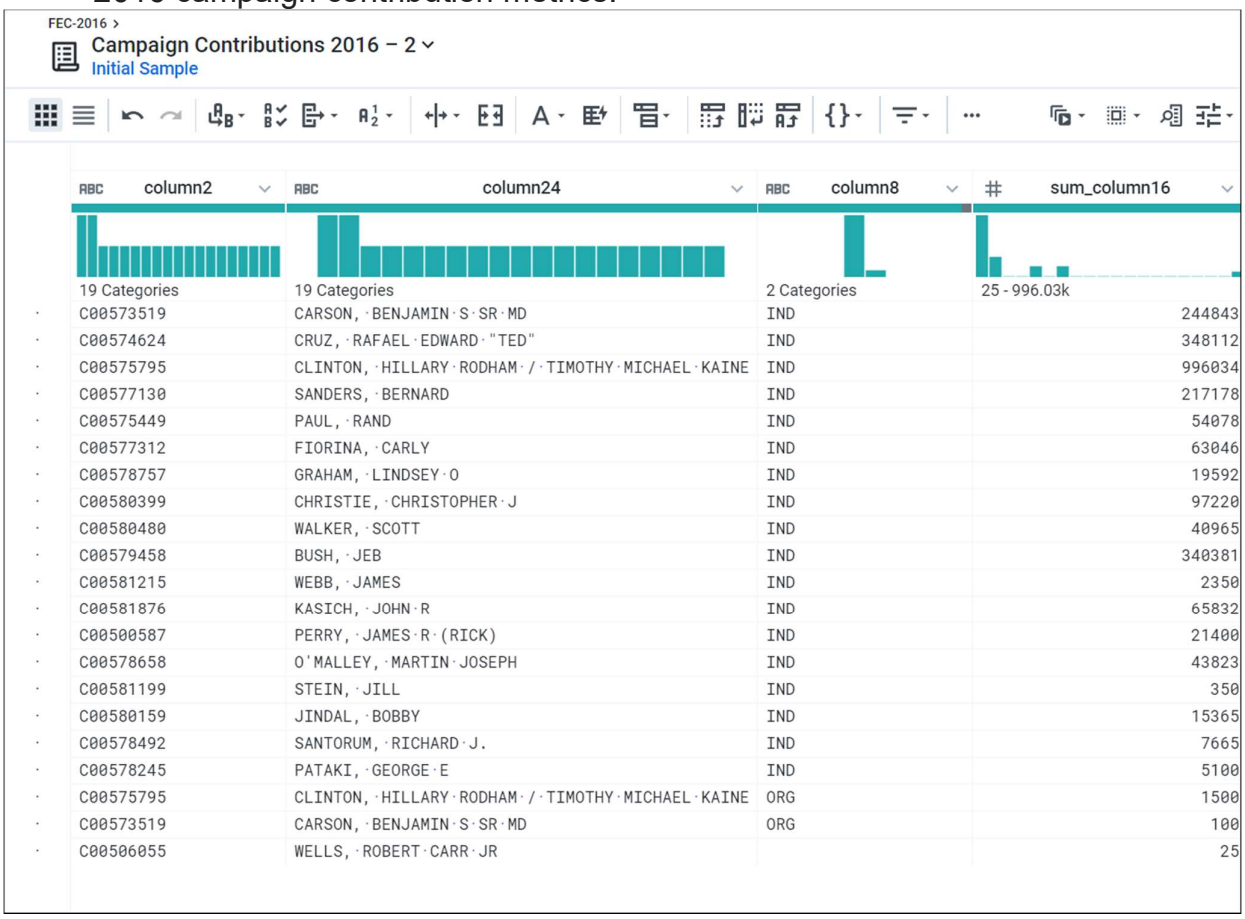
1. At the top of the Recipe panel on the left, click on **New Step** and enter the following formula in the **Transformation** search box to preview the aggregated data.

```
pivot value:sum(column16),average(column16),countif(column16 > 0) group:
column2,column24,column8
```

An initial sample of the joined and aggregated data is displayed, representing a summary table of US presidential candidates and their 2016 campaign contribution metrics.



2. Click **Add** to open a summary table of major US presidential candidates and their 2016 campaign contribution metrics.



column2	column24	column8	sum_column16
19 Categories	19 Categories	2 Categories	25 - 996.03k
C00573519	CARSON, BENJAMIN S SR MD	IND	244843
C00574624	CRUZ, RAFAEL EDWARD "TED"	IND	348112
C00575795	CLINTON, HILLARY RODHAM / TIMOTHY MICHAEL KAINE	IND	996034
C00577130	SANDERS, BERNARD	IND	217178
C00575449	PAUL, RAND	IND	54078
C00577312	FIORINA, CARLY	IND	63046
C00578757	GRAHAM, LINDSEY O	IND	19592
C00580399	CHRISTIE, CHRISTOPHER J	IND	97220
C00580480	WALKER, SCOTT	IND	40965
C00579458	BUSH, JEB	IND	340381
C00581215	WEBB, JAMES	IND	2350
C00581876	KASICH, JOHN R	IND	65832
C00500587	PERRY, JAMES R (RICK)	IND	21400
C00578658	O'MALLEY, MARTIN JOSEPH	IND	43823
C00581199	STEIN, JILL	IND	350
C00580159	JINDAL, BOBBY	IND	15365
C00578492	SANTORUM, RICHARD J.	IND	7665
C00578245	PATAKI, GEORGE E	IND	5100
C00575795	CLINTON, HILLARY RODHAM / TIMOTHY MICHAEL KAINE	ORG	1500
C00573519	CARSON, BENJAMIN S SR MD	ORG	100
C00506055	WELLS, ROBERT CARR JR		25

Rename columns

You can make the data easier to interpret by renaming the columns. Add each of the renaming and rounding steps individually to the recipe by clicking **New Step**, then enter:

```
rename type: manual mapping: [column24,'Candidate_Name'],
[column2,'Candidate_ID'],[column8,'Party_Affiliation'],
[sum_column16,'Total_Contribution_Sum'], [average_column16,'Average_Contribution_Sum'],
[countif,'Number_of_Contributions']
```

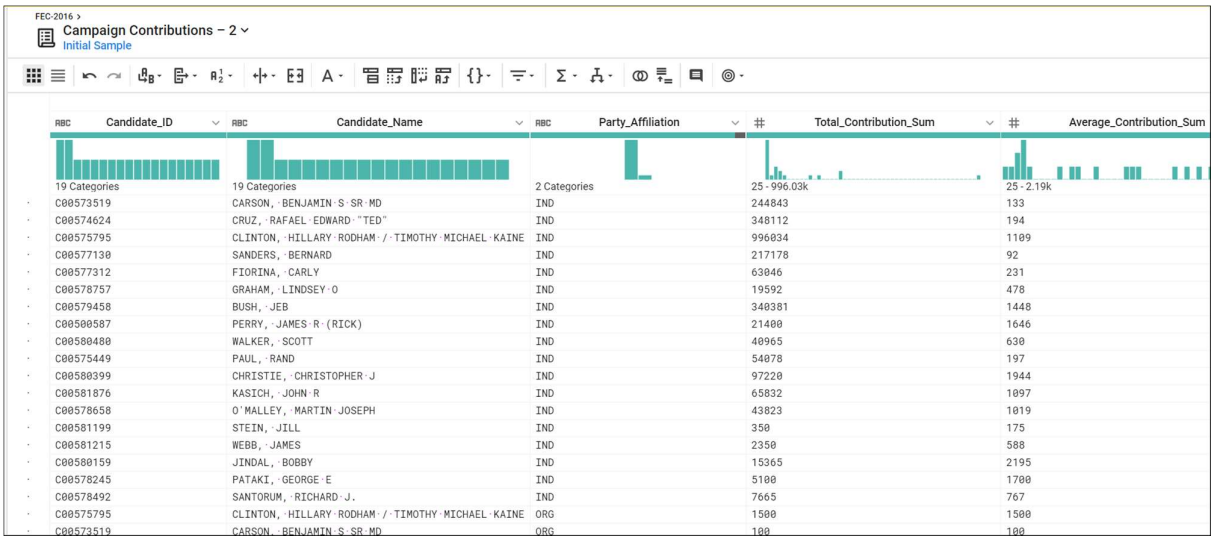
Then click **Add**.

Add in this last **New Step** to round the Average Contribution amount:

```
set col: Average_Contribution_Sum value: round(Average_Contribution_Sum)
```

Then click **Add**.

Your results look something like this:



Candidate_ID	Candidate_Name	Party_Affiliation	Total_Contribution_Sum	Average_Contribution_Sum
19 Categories	19 Categories	2 Categories	25 - 996.03k	25 - 2.19k
C00573519	CARSON, BENJAMIN S SR MD	IND	244843	133
C00574624	CRUZ, RAFAEL EDWARD "TED"	IND	348112	194
C00575795	CLINTON, HILLARY RODHAM / TIMOTHY MICHAEL KAINE	IND	996034	1109
C00577130	SANDERS, BERNARD	IND	217178	92
C00577312	FIORINA, CARLY	IND	63046	231
C00578757	GRAHAM, LINDSEY O	IND	19592	478
C00579458	BUSH, JEB	IND	340381	1448
C00500587	PERRY, JAMES R (RICK)	IND	21400	1646
C00580480	WALKER, SCOTT	IND	40965	630
C00575449	PAUL, RAND	IND	54078	197
C00580399	CHRISTIE, CHRISTOPHER J	IND	97220	1944
C00581876	KASICH, JOHN R	IND	65832	1097
C00578658	O'MALLEY, MARTIN JOSEPH	IND	43823	1019
C00581199	STEIN, JILL	IND	350	175
C00581215	WEBB, JAMES	IND	2350	588
C00580159	JINDAL, BOBBY	IND	15365	2195
C00578245	PATAKI, GEORGE E	IND	5100	1700
C00578492	SANTORUM, RICHARD J.	IND	7665	767
C00575795	CLINTON, HILLARY RODHAM / TIMOTHY MICHAEL KAINE	ORG	1500	1500
C00573519	CARSON, BENJAMIN S SR MD	ORG	100	100

Congratulations!

You used Dataprep to add a dataset and created recipes to wrangle the data into meaningful results.

Next Steps / Learn More

This lab is part of a series of labs called Qwik Starts. These labs are designed to give you a little taste of the many features available with Google Cloud. Search for "Qwik Starts" in the lab catalog to find the next lab you'd like to take!

Google Cloud Training & Certification

...helps you make the most of Google Cloud technologies. Our classes include technical skills and best practices to help you get up to speed quickly and continue your learning journey. We offer fundamental to advanced level training, with on-demand, live, and virtual options to suit your busy schedule. Certifications help you validate and prove your skill and expertise in Google Cloud technologies.

Manual Last Updated March 19, 2021
Lab Last Tested March 19, 2021