

Dataprep: Qwik Start | Google Cloud Skills Boost

Qwiklabs : 12-15 minutes

This lab was developed with our partner, [Trifacta](#). Your personal information may be shared with Trifacta, the lab sponsor, if you have opted-in to receive product updates, announcements, and offers in your Account Profile.

GSP105



Google Cloud Self-Paced Labs

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 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.

To complete this lab, you need:

- Access to a standard internet browser (Chrome browser recommended).

Note: Use an Incognito or private browser window to run this lab. This prevents any conflicts between your personal account and the Student account, which may cause extra charges incurred to your personal account.

- Time to complete the lab---remember, once you start, you cannot pause a lab.

Note: If you already have your own personal Google Cloud account or project, do not use it for this lab to avoid extra charges to your account.

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 the **Lab Details** panel with the following:

- The **Open Google Console** button
- Time remaining
- The temporary credentials that you must use for this lab
- Other information, if needed, to step through this lab

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

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

Note: If you see the **Choose an account** dialog, click **Use Another Account**.

3. If necessary, copy the **Username** from the **Lab Details** panel and paste it into the **Sign in** dialog. Click **Next**.

4. Copy the **Password** from the **Lab Details** panel and paste it into the **Welcome** dialog. Click **Next**.

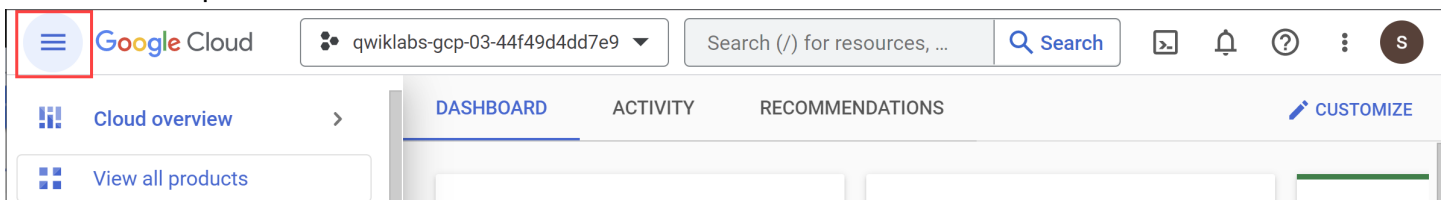
Important: You must use the credentials from the left panel. Do not use your Google Cloud Skills Boost credentials. **Note:** Using your own Google Cloud account for this lab may incur extra charges.

5. 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.



Task 1. Create a Cloud Storage bucket in your project

1. In the Cloud Console, select **Navigation menu**(≡) > **Cloud Storage** > **Buckets**.

2. Click **Create bucket**.

3. In the **Create a bucket** dialog, **Name** the bucket a unique name. Leave other settings at their default value.

Note: Learn more about naming buckets from [Bucket naming guidelines](#).

4. Uncheck **Enforce public access prevention on this bucket** for Choose how to control access to objects.
5. 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 a Cloud Storage bucket, you see an assessment score.

Create a Cloud Storage bucket

Task 2. 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 student username to sign in to Cloud Dataprep by Trifacta. Your username is the **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 setup** screen to create the default storage location.

Dataprep opens.

Test completed task

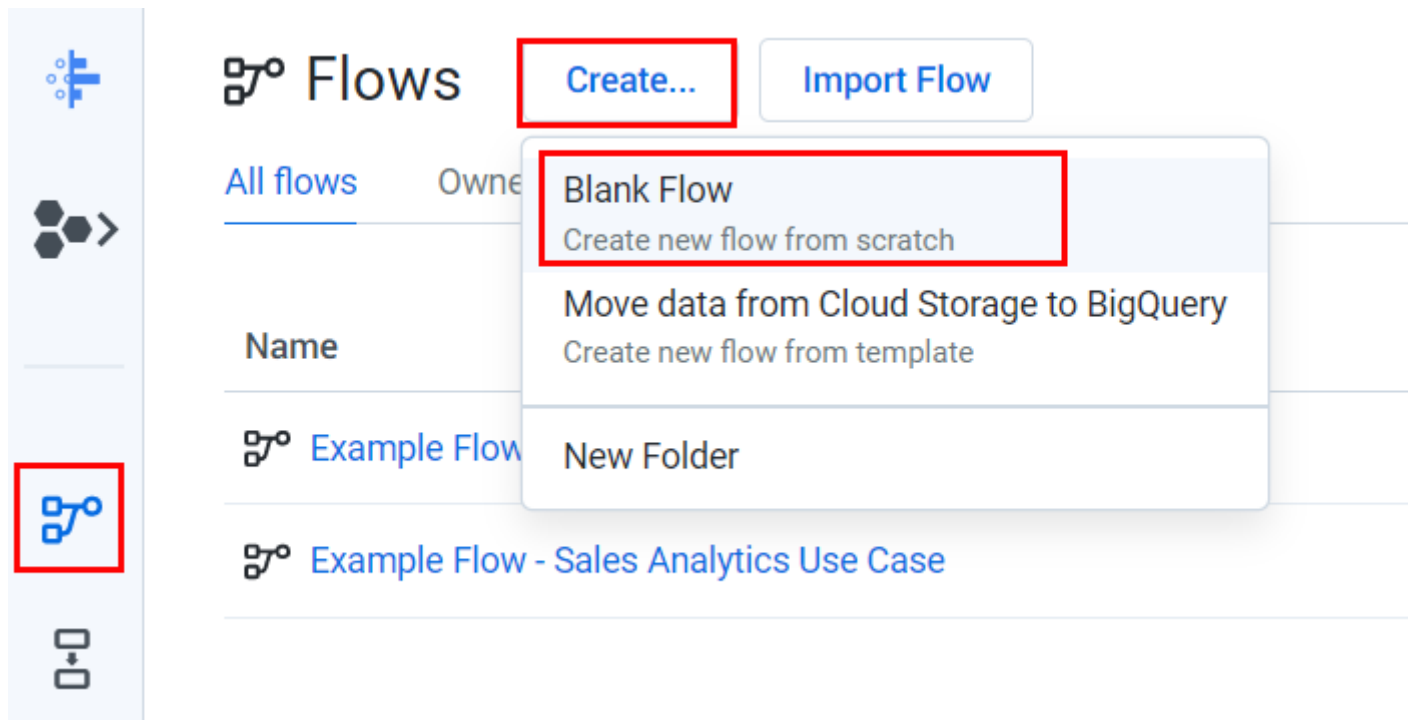
Click **Check my progress** to verify your performed task. If you have successfully initialized Cloud Dataprep with default storage location, you see an assessment score.

Initialize Cloud Dataprep

Task 3. Create a flow

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

1. Click **Flows** icon, then the **Create** button, then select **Blank Flow** :



2. Click on **Untitled Flow**, then 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 then describe the flow as "United States Federal Elections Commission 2016".
3. Click **OK**.

The FEC-2016 flow page opens.

Task 4. Import datasets


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

1. Click **Add Datasets**, then select the **Import Datasets** link.
2. In the left menu pane, select **Cloud Storage** to import datasets from Cloud Storage, then click on the pencil to edit the file path.



Import Data and Add to Flow

Search...

Choose a file or folder

Cloud Storage 

Search...

NAME	SIZE
 dataprep-staging-51d7a1ba-e0c5-48e...	
 qwiklabs-gcp-00-46cd4322783d	

3. Type `gs://splis/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.

4. Click **us-fec/**.

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

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

7. Both datasets are listed in the right pane; click **Import & Add to Flow**.

2 New Datasets

Clear All

Campaign Contributions 2016

X

Add a Description

ABC column2	ABC column3	ABC column4
C00000935	A	M10
C00000935	A	M4
C00000935	A	M6
C00000935	A	M7
C00000935	A	M8

Edit settings

Candidate Master 2016

X

Add a Description

ABC column2	ABC column3
H0AK00097	COX, JOHN R.
H0AL02087	ROBY, MARTHA
H0AL02095	JOHN, ROBERT E JR
H0AL05049	CRAMER, ROBERT E
H0AL05163	BROOKS, MO

Edit settings

Import & Add to Flow

Cancel

You see both datasets listed as a flow.

Task 5. Prep the candidate file

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

Dataset

Candidate Master 2016

Recipe

+

Output

Candidate Master 2016

Dataset

Campaign Contributions 2016

Details

X

Candidate Master 2016

Edit Recipe

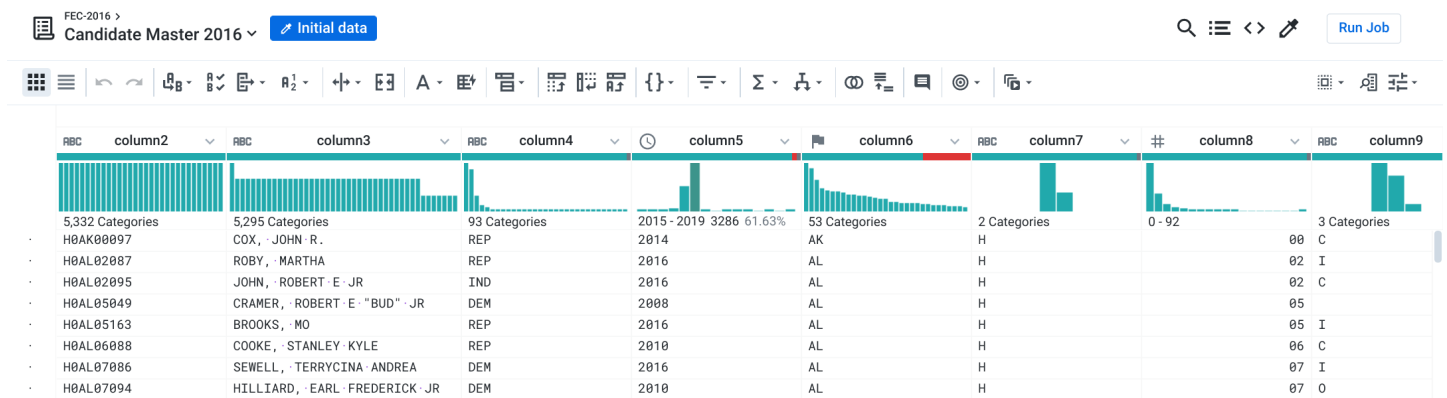
Add

Recipe

Data

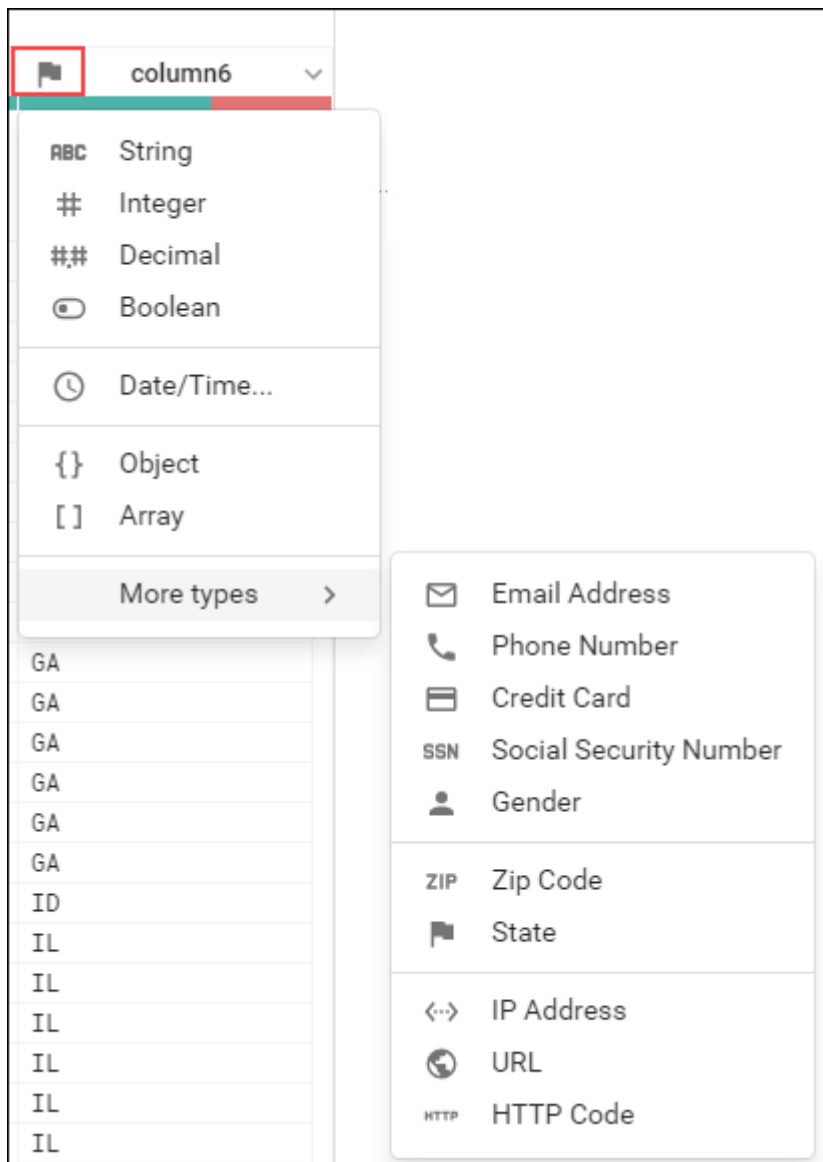
Steps Preview

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. To see data types, click the column icon:

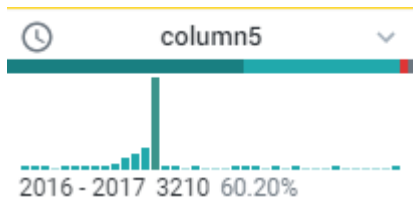


- Notice also that when you click the name of the column, a **Details** panel opens on the right.

- 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.

1. 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 the year 2016.



This creates a step where these values are selected.

2. In the **Suggestions** panel on the right, in the **Keep rows** section, click **Add** to add this step to your recipe.

Suggestions ×

Keep rows

where (DATE(2016, 1, 1) <= column5) && (column5 < DATE(2018, 1, 1))

Edit
Add

Delete rows

where (DATE(2016, 1, 1) <= column5) && (column5 < DATE(2018, 1, 1))

Set

Set column5 to IF((DATE(2016, 1, 1) <= column5) && (column5 < DATE(2018, 1, 1)), NULL(), \$col)

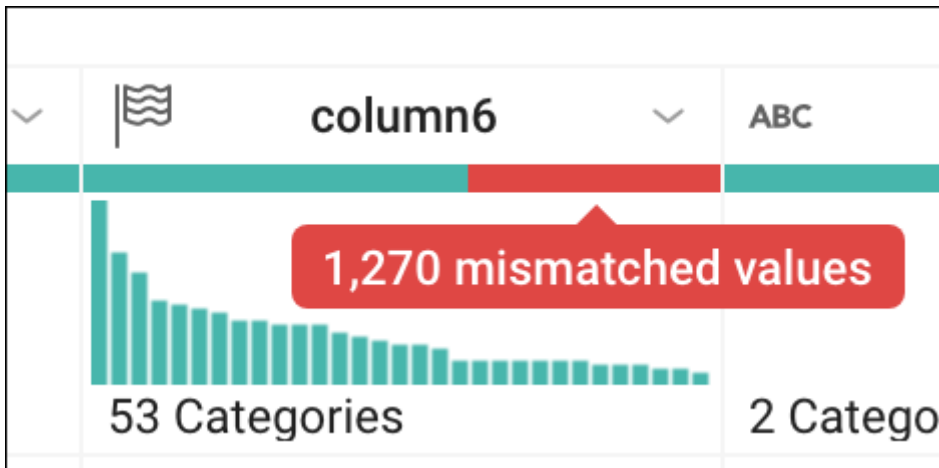
Create a new column

(DATE(2016, 1, 1) <= column5) && (column5 < DATE(2018, 1, 1))

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

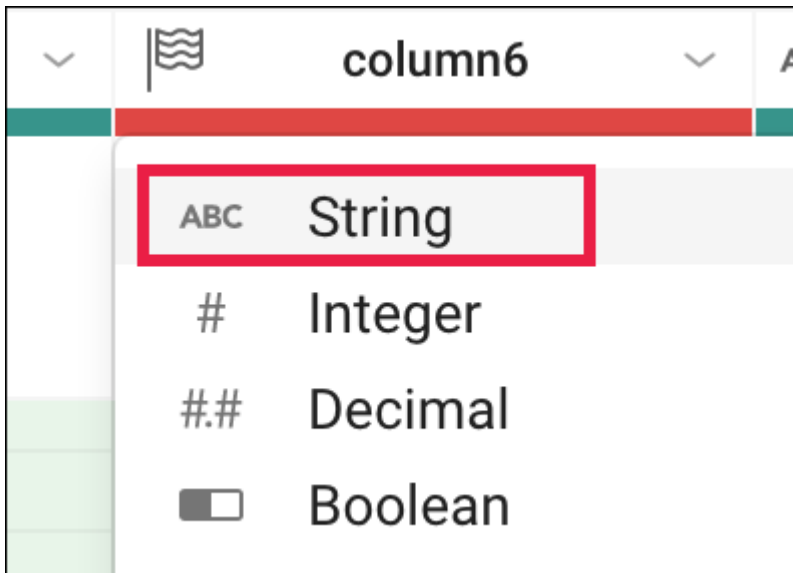
Keep rows where(DATE(2016, 1, 1) <= column5) && (column5 < DATE(2018, 1, 1))

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



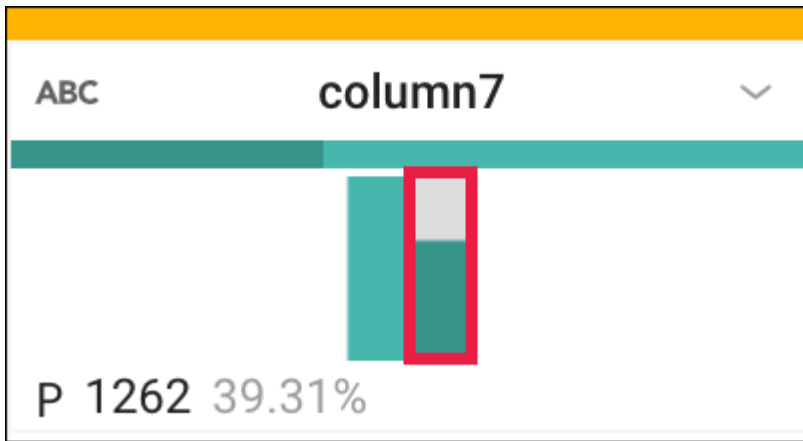
Scroll down to the bottom (highlighted in red) find the mismatched values and notice how most of these records have the value "P" in column7, and "US" in column6. The mismatch occurs because column6 is marked as a "State" column (indicated by the flag icon), but there are non-state (such as "US") values.

4. 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.

5. 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.



The screenshot shows a data table with a header row containing 'ABC' and 'column7'. Below the header, there is a row with a teal bar and a cell in the 'column7' column that is highlighted with a red box. At the bottom of the table, the text 'P 1262 39.31%' is visible.

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

Keep rows

where column7 == 'P'

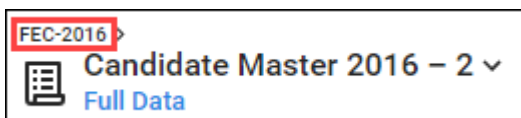
[Edit](#)[Add](#)

Task 6. Wrangle the Contributions file and join it to the Candidates 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.

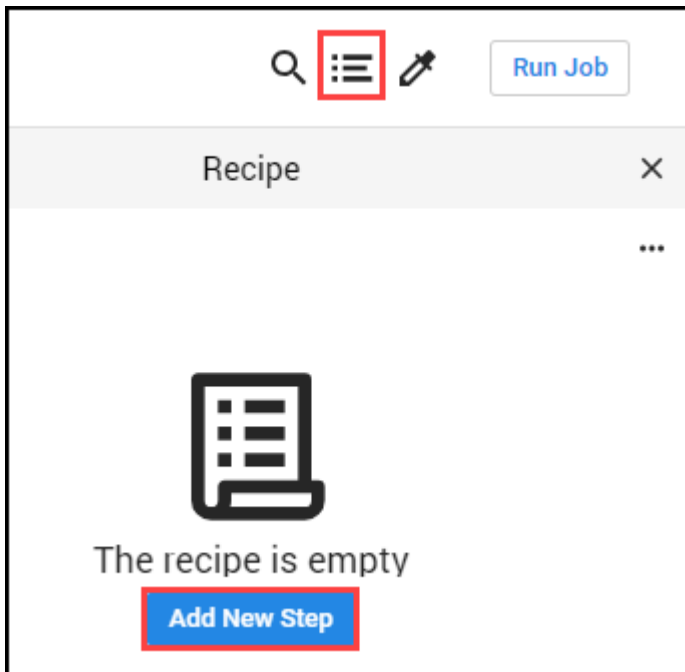
1. Click on **FEC-2016** (the dataset selector) at the top of the grid view page.



2. Click to select the grayed out **Campaign Contributions 2016**.

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.

Column required

All ▼

Find

``{start}`|`{end}``

Replace with required

String

[Advanced options](#) ▼

Cancel Add

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.

[< Recipe](#) Search transformations [×](#)

Join [▼](#)

Join datasets

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

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

✓	Candidate Master 2016	Today at 4:40 PM	FEC-2016
---	-----------------------	------------------	----------

10. On the right side, hover in the Join keys section, then click on the pencil (Edit icon).

[< Joined-in Data](#) Join Conditions [×](#)

Join type required

Inner [▼](#)

Join keys [?](#) [Add](#)

ABC column9

= (Equal to)

ABC column3

0% match

Results summary

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**.

< Join Conditions

Add Key

×

Current

required

ABC column9

×

▼

Joined-in

required

ABC column3

×

▼

☐ Fuzzy match

☐ Ignore case

☐ Ignore special characters

☐ Ignore whitespace

Suggested join keys ?

ABC column9 = ABC column3

ABC column10 = ABC column14

ABC column2 = ABC column11

ABC column2 = ABC column2

ABC column13 = ABC column3

ABC column17 = ABC column2

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 "Column" label to add all columns of both datasets to the joined dataset.

All (36)

Current (21)

Joined-In (15)

<div><input checked="" type="checkbox"/></div>	Column	Source
<div><input type="checkbox"/></div> <div></div>	column2	<div><div></div><div></div></div>
<div><input type="checkbox"/></div> <div></div>	column11	<div><div></div><div></div></div>
<div><input type="checkbox"/></div>	column3	<div><div></div><div></div></div>
<div><input type="checkbox"/></div>	column4	<div><div></div><div></div></div>
<div><input type="checkbox"/></div>	column5	<div><div></div><div></div></div>
<div><input type="checkbox"/></div>	column6	<div><div></div><div></div></div>
<div><input type="checkbox"/></div>	column7	<div><div></div><div></div></div>
<div><input type="checkbox"/></div>	column8	<div><div></div><div></div></div>
<div><input type="checkbox"/></div>	column9	<div><div></div><div></div></div>

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

Task 7. 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 right, 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.

Task 8. Rename columns

You can make the data easier to interpret by renaming the columns.

1. 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']

2. Then click **Add**.
3. Add in this last **New Step** to round the Average Contribution amount:

set col: Average_Contribution_Sum value: round(Average_Contribution_Sum)

4. Then click **Add**.

Your results look something like this:

ABC	Candidate_ID	ABC	Candidate_Name	ABC	Party_Affiliation	#	Total_Contribution_Sum
	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	

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 and 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 Feb 1, 2022

Lab Last Tested Feb 1, 2022

Copyright 2023 Google LLC All rights reserved. Google and the Google logo are trademarks of Google LLC. All other company and product names may be trademarks of the respective companies with which they are associated.