Dataprep: Qwik Start

Activate Cloud Shell

Cloud Shell is a virtual machine that is loaded with development tools. It offers a persistent 5GB home directory and runs on the Google Cloud. Cloud Shell provides command-line access to your Google Cloud resources.

1. Click **Activate Cloud Shell** at the top of the Google Cloud console.

2. Click through the following windows:

   - Continue through the Cloud Shell information window.
   - Authorize Cloud Shell to use your credentials to make Google Cloud API calls.

When you are connected, you are already authenticated, and the project is set to your **Project_ID**, `qwiklabs-gcp-02-2107941e3abf`. The output contains a line that declares the **Project_ID** for this session:

```
Your Cloud Platform project in this session is set to qwiklabs-gcp-02-2107941e3abf
```

`gcloud` is the command-line tool for Google Cloud. It comes pre-installed on Cloud Shell and supports tab-completion.

3. (Optional) You can list the active account name with this command:

```
gcloud auth list
```

Copied!

content_copy

4. Click **Authorize**.

Output:

```
ACTIVE: *
ACCOUNT: student-01-0cfc3c5ffa62@qwiklabs.net

To set the active account, run:
    $ gcloud config set account `ACCOUNT`
```

5. (Optional) You can list the project ID with this command:

```
gcloud config list project
```

Copied!

content_copy

Output:

```
[core]
project = qwiklabs-gcp-02-2107941e3abf
```

Note: For full documentation of `gcloud`, in Google Cloud, refer to [the gcloud CLI overview guide](#).

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.

Task 2. Initialize Cloud Dataprep

1. Open **Cloud Shell** and run the following command:

```
gcloud beta services identity create --service=dataprep.googleapis.com
```

Copied!

content_copy

You should see a message saying the service identity was created.

2. Select **Navigation menu** > **Dataprep**.

3. Check to accept the Google Dataprep Terms of Service, then click **Accept**.

4. Check to authorize sharing your account information with Trifacta, then click **Agree and Continue**.

5. Click **Allow** to allow Trifacta to access project data.

6. Click your student username to sign in to Cloud Dataprep by Trifacta. Your username is the **Username** in the left panel in your lab.

7. Click **Allow** to grant Cloud Dataprep access to your Google Cloud lab account.

8. Check to agree to Trifacta Terms of Service, and then click **Accept**.

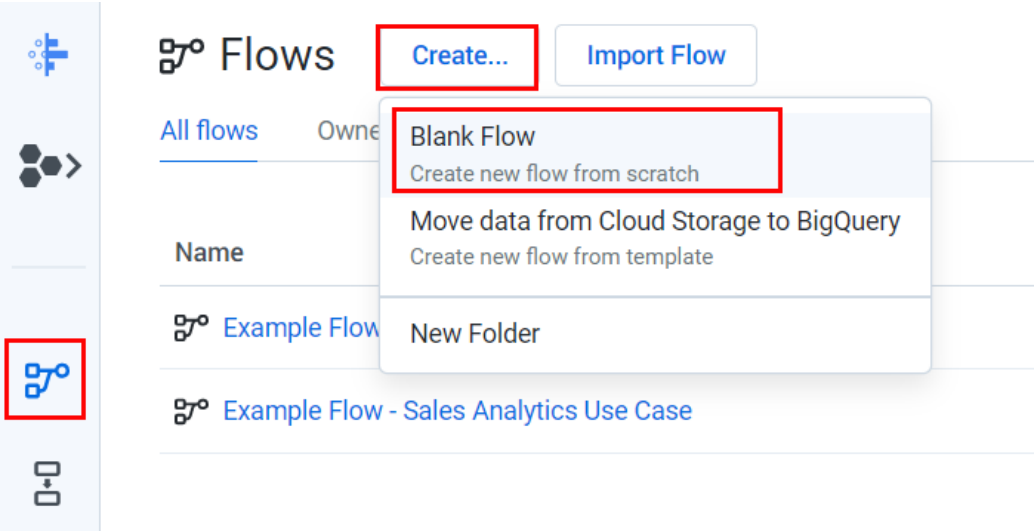
9. Click **Continue** on the **First time setup** screen to create the default storage location.

Dataprep opens.

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**.

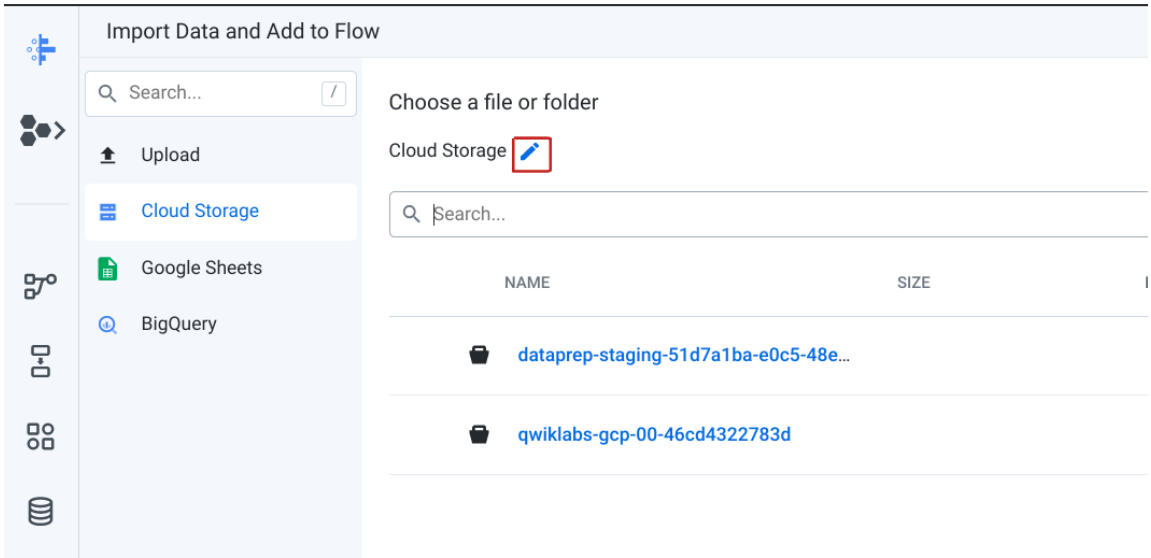
The FEC-2016 flow page opens.

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.



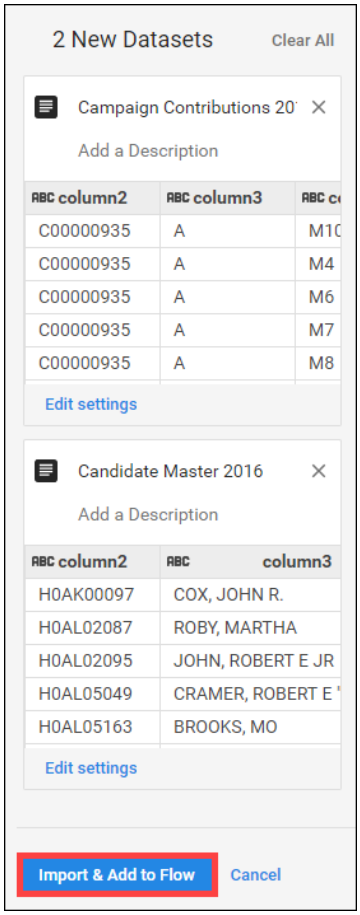
3. Type `gs://spls/gsp105` in the **Choose a file or folder** text box, then click **Go**. You may have to widen the browser window to see the **Go** and **Cancel** buttons.

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**.

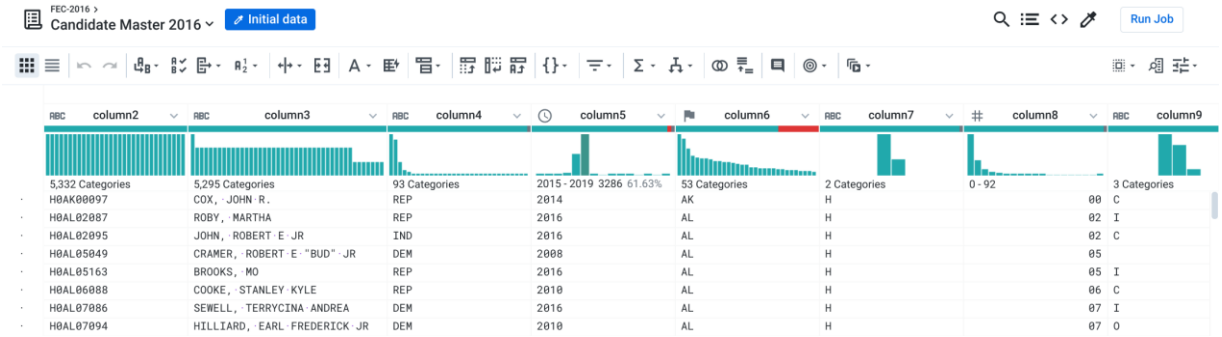


You see both datasets listed as a flow.

Task 5. Prep the candidate file

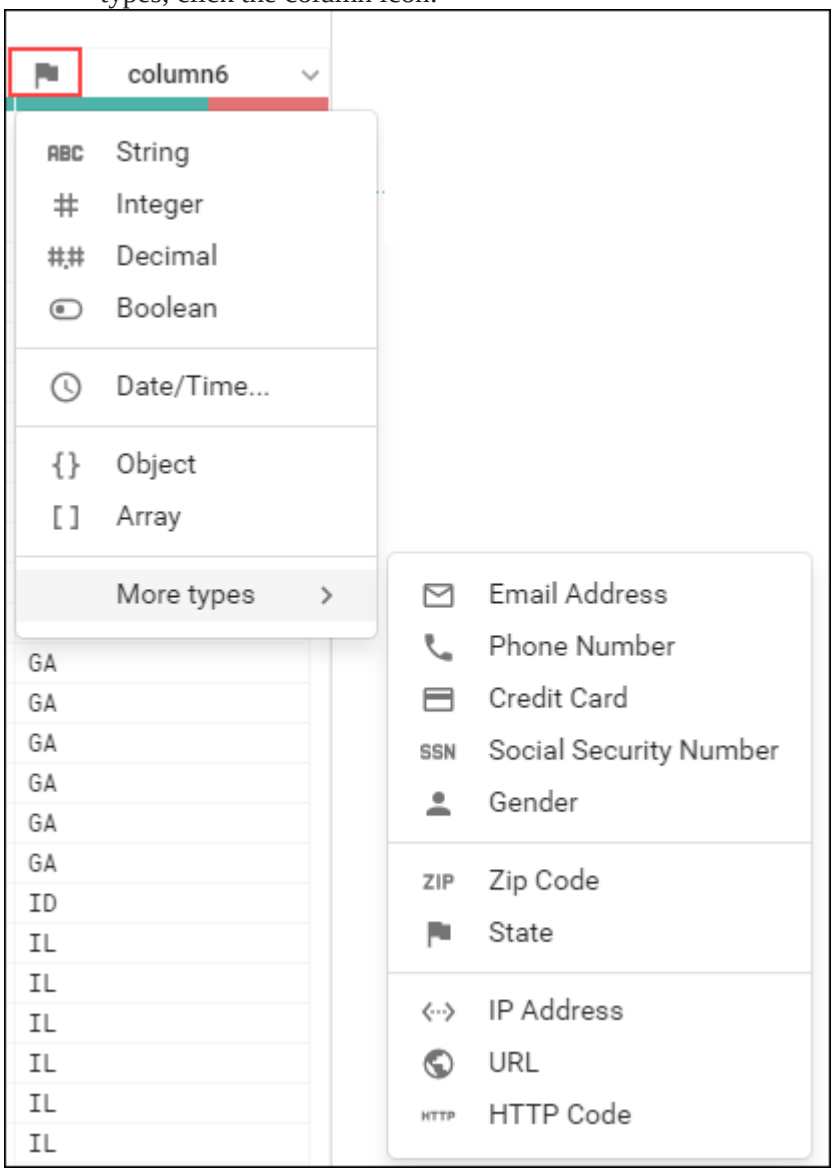
1. By default, the Candidate Master 2016 dataset is selected. In the right pane, click **Edit Recipe**.

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.

2. Each of the column heads have a Name and value that specify the data type. To see data types, click the column icon:

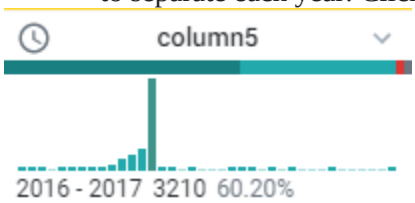


3. Notice also that when you click the name of the column, a **Details** panel opens on the right.

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

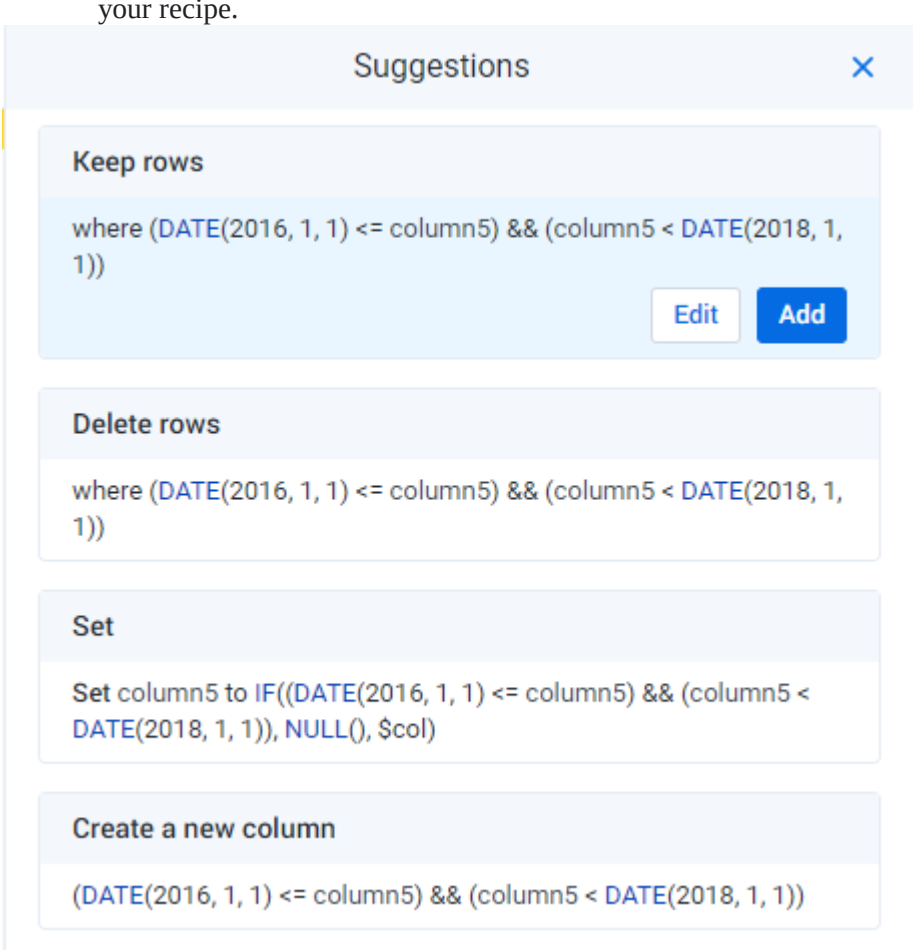
In the following steps you explore data in the grid view and apply transformation steps to your recipe.

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.

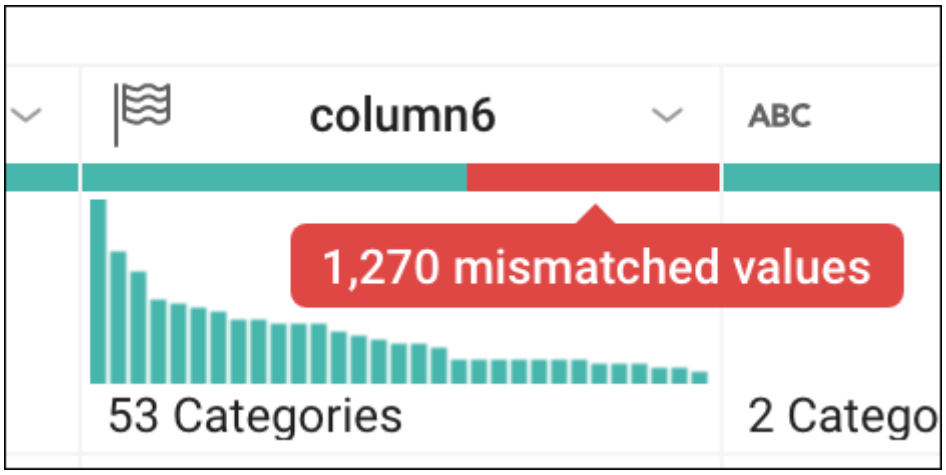
2. In the **Suggestions** panel on the right, in the **Keep rows** section, click **Add** to add this step to your recipe.



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

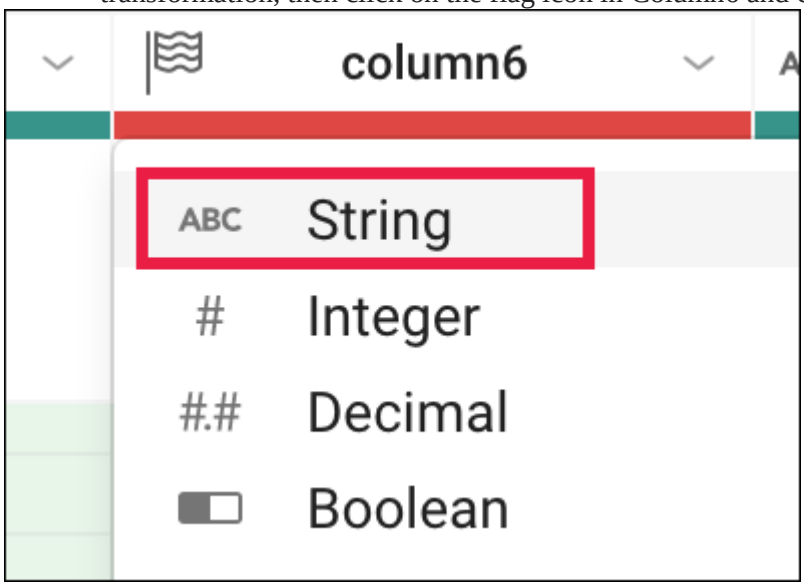
```
Keep rows where(DATE(2016, 1, 1) <= column5) && (column5 <
DATE(2018, 1, 1))
```

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



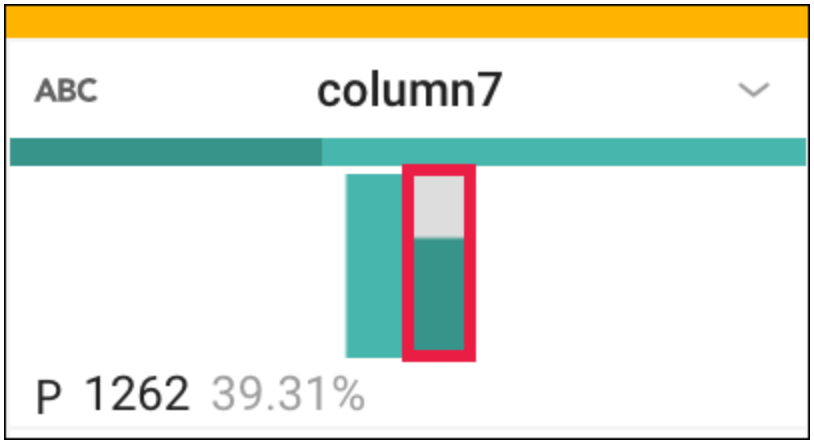
Scroll down to the bottom (highlighted in red) find the mismatched values and notice how most of these records have the value "P" in column7, and "US" in column6. The mismatch occurs because column6 is marked as a "State" column (indicated by the flag icon), but there are non-state (such as "US") values.

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.

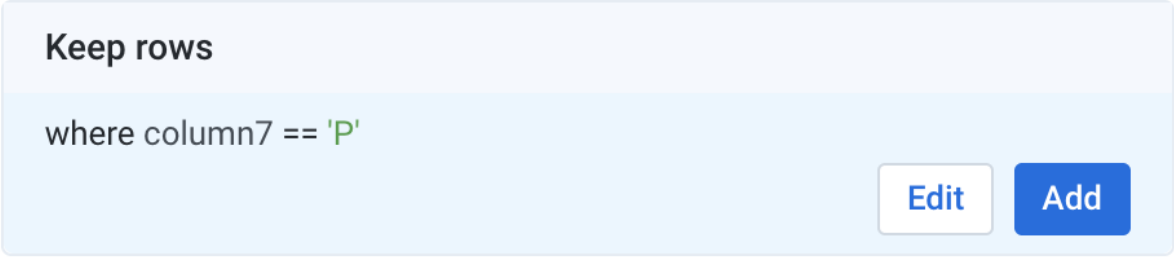


There is no longer a mismatch and the column marker is now green.

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.



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

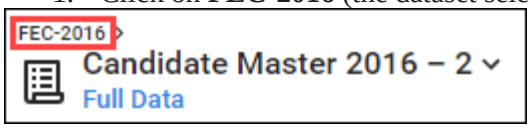


Task 6. Wrangle the Contributions file and join it to the Candidates file

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.

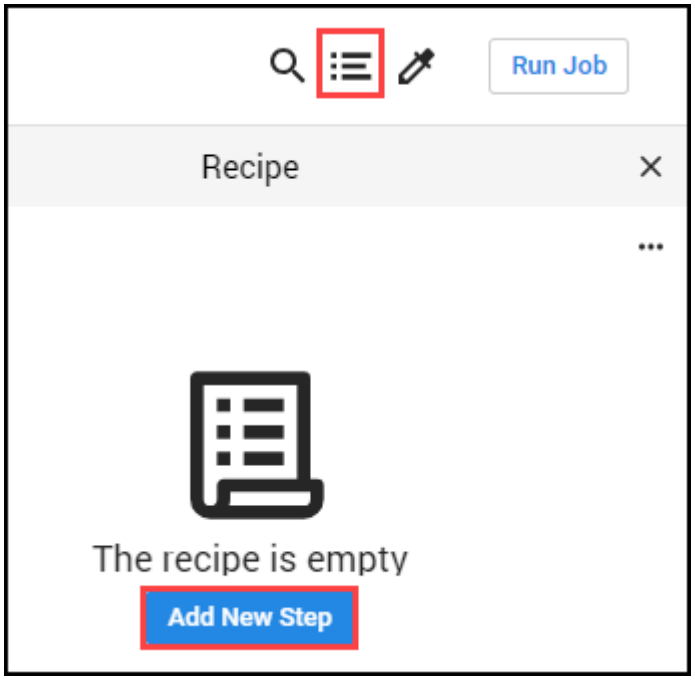
1. Click on **FEC-2016** (the dataset selector) at the top of the grid view page.



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.

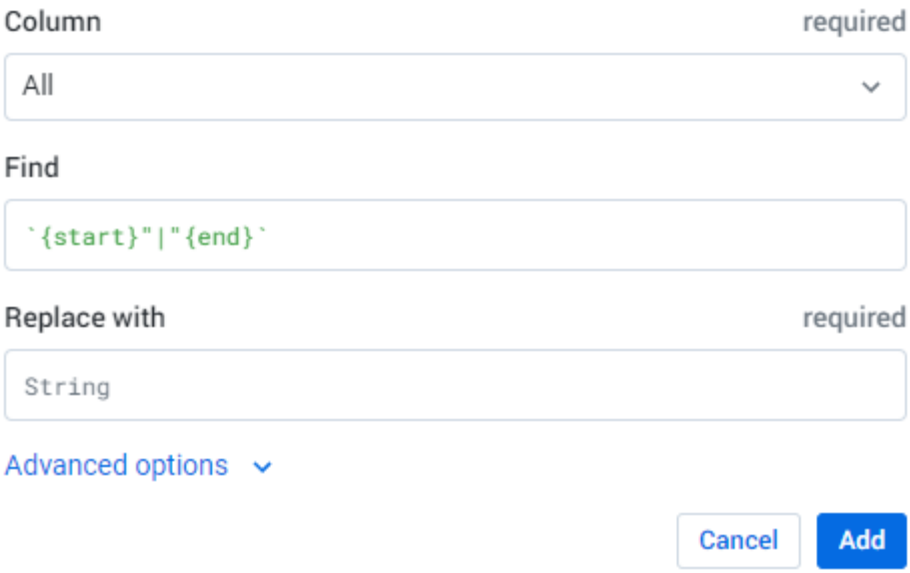
5. Insert the following Wrangle language command in the Search box:

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

Copied!

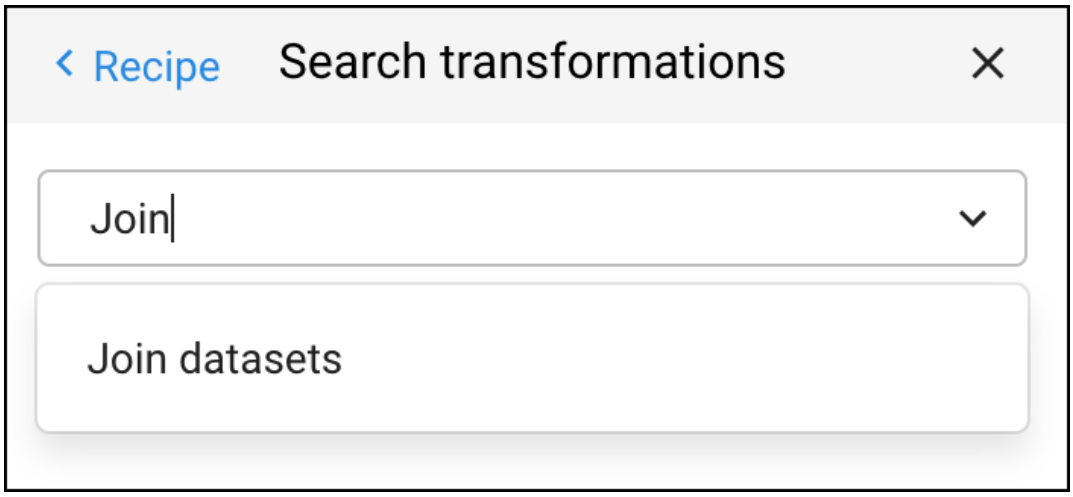
content_copy

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



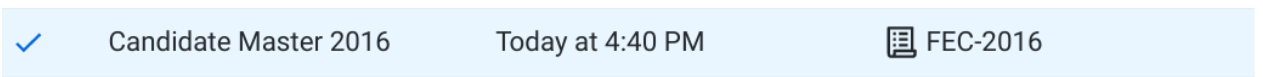
6. Click **Add** to add the transform to the recipe.

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

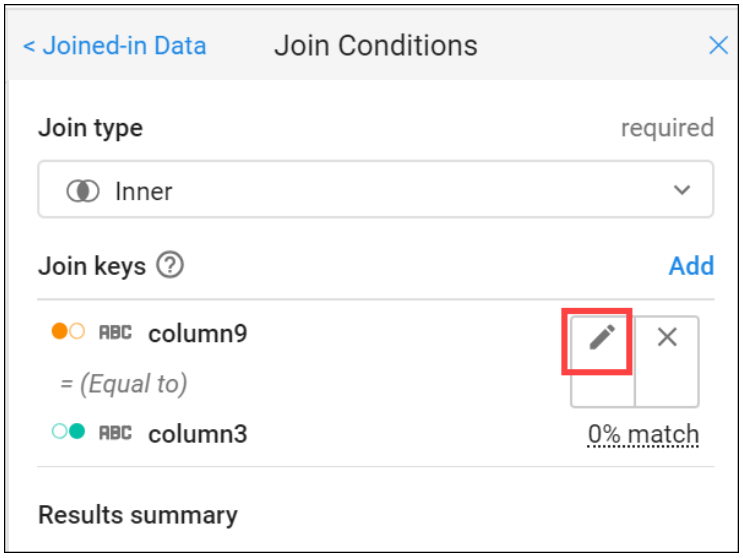


8. Click **Join datasets** to open the Joins page.

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

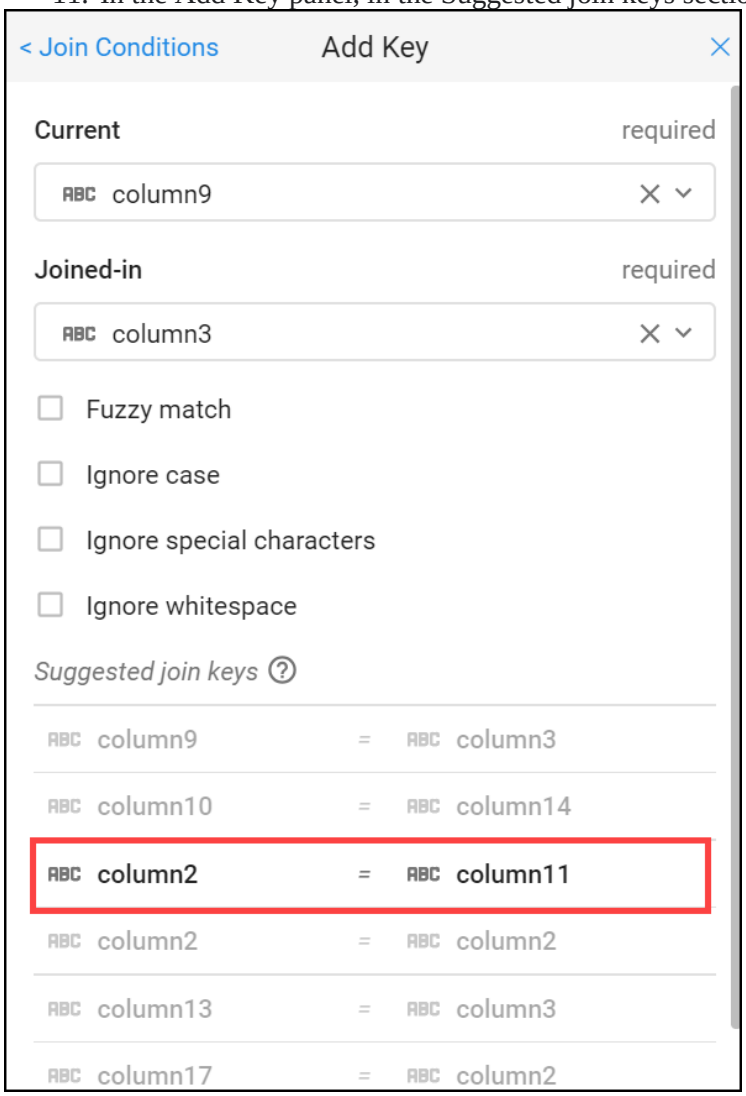


10. On the right side, hover in the Join keys section, then click on the pencil (Edit icon).



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

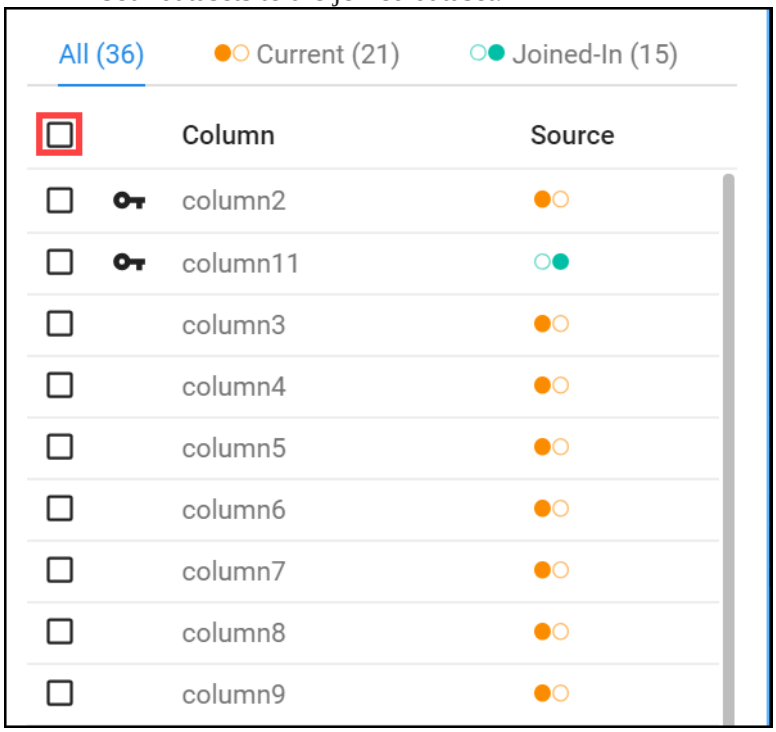
11. In the Add Key panel, in the Suggested join keys section, click **column2 = column11**.



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 "Column" 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.

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.

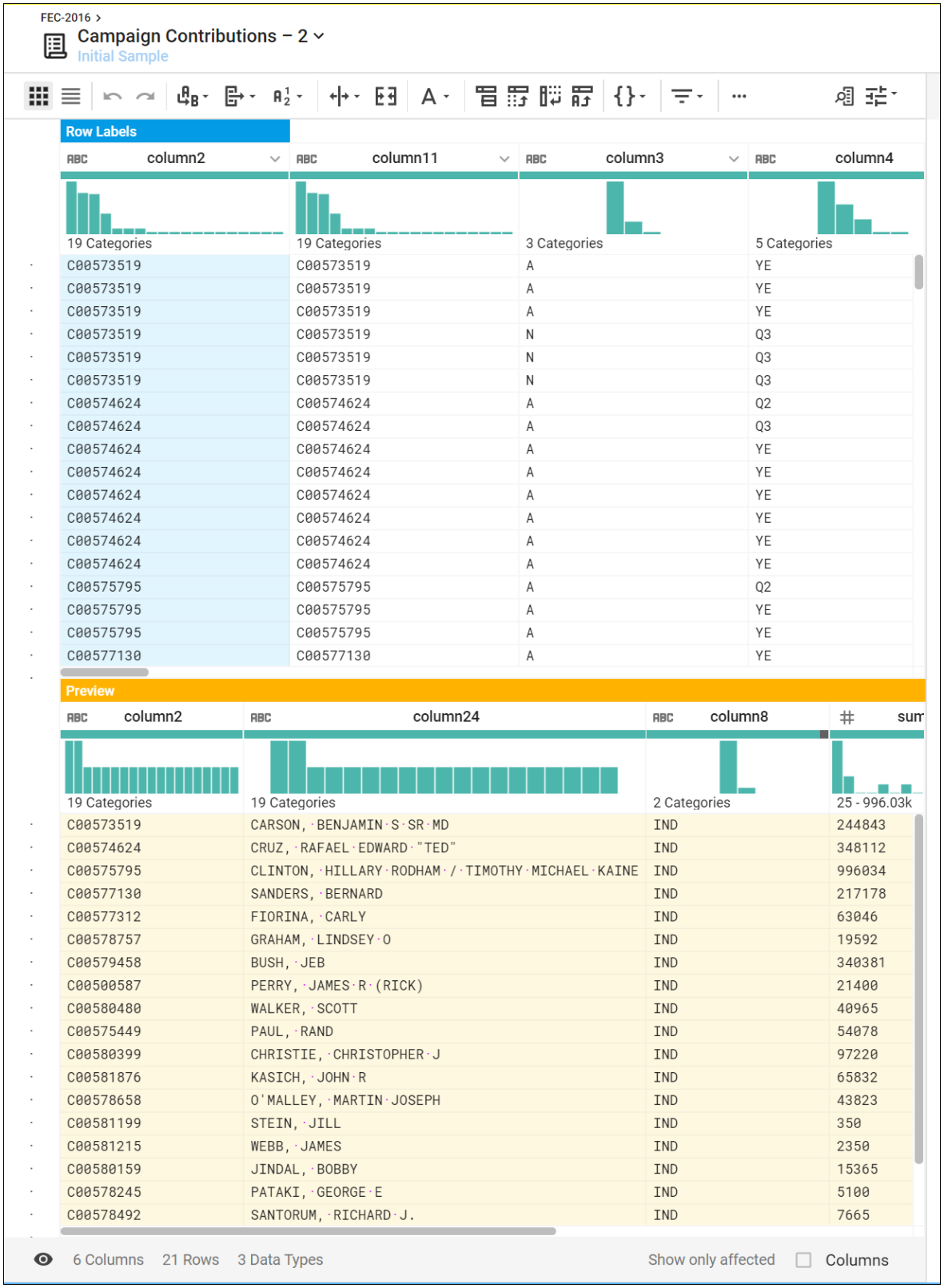
1. At the top of the Recipe panel on the right, click on **New Step** and enter the following formula in the **Transformation** search box to preview the aggregated data.

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

Copied!

content_copy

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']
```

Copied!

content_copy

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)
```

Copied!

content_copy

4. Then click **Add**.

Your results look something like this:

Candidate_ID	Candidate_Name	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