Dataprep: Qwik Start

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.
4. Uncheck **Enforce public access prevention on this bucket** for Choose how to control access to objects.
5. Click **Create**.

Task 2. Initialize Cloud Dataprep

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

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

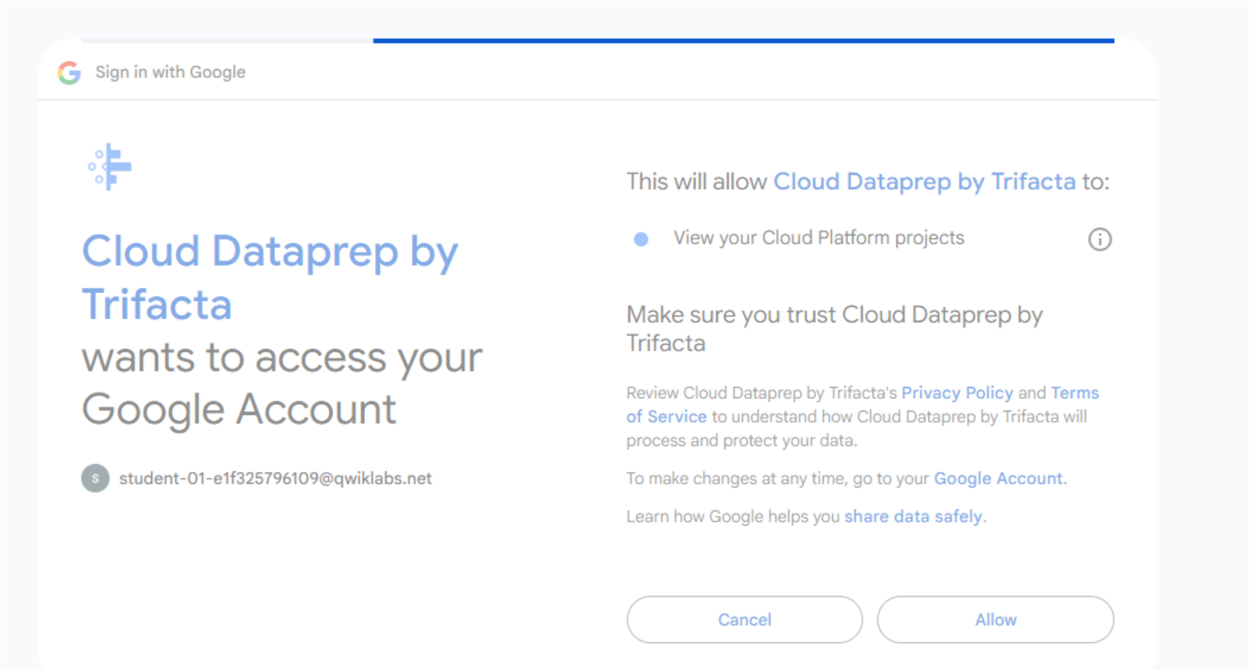
2. In the Cloud console, go to the **Navigation menu**, click **View All Products** and under **Analytics** select **Alteryx Designer Cloud**.
3. Check to accept the Google Dataprep Terms of Service, then click **Accept**.
4. Check to authorize sharing your account information with Trifacta, then click **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.

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.



Sign in with Google

Cloud Dataprep by Trifacta wants to access your Google Account

student-01-e1f325796109@qwiklabs.net

This will allow Cloud Dataprep by Trifacta to:

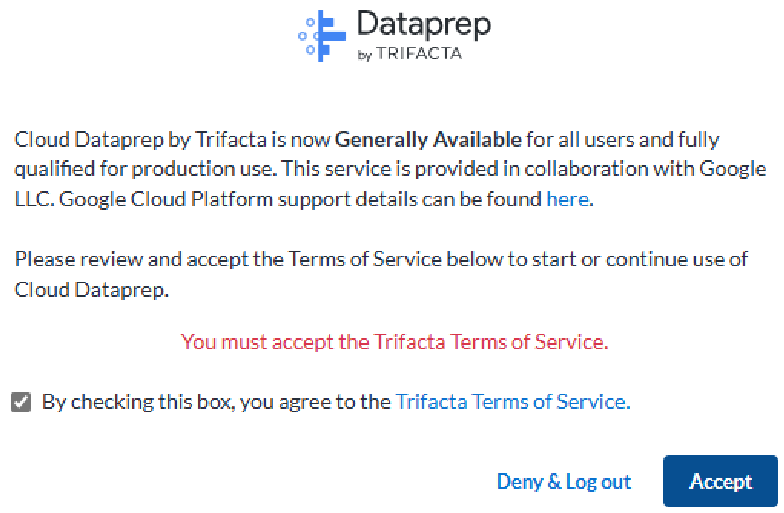
- View your Cloud Platform projects

Make sure you trust Cloud Dataprep by Trifacta

Review Cloud Dataprep by Trifacta's Privacy Policy and Terms of Service to understand how Cloud Dataprep by Trifacta will process and protect your data.

To make changes at any time, go to your Google Account.

Learn how Google helps you share data safely.



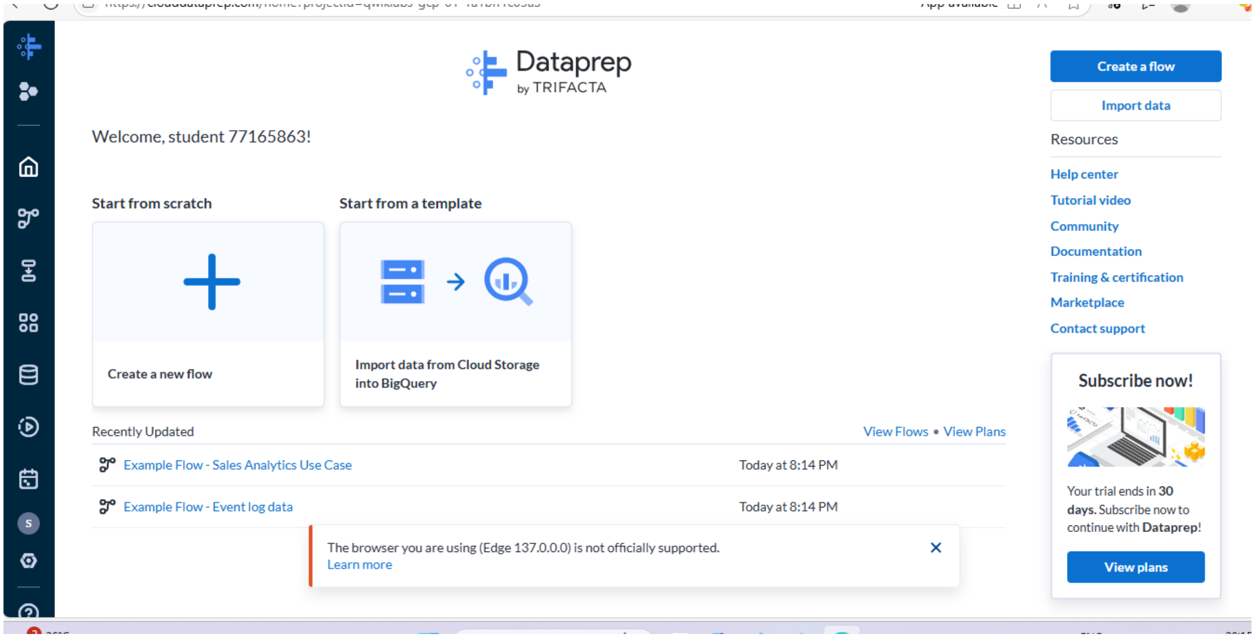
Dataprep by TRIFACTA

Cloud Dataprep by Trifacta is now **Generally Available** for all users and fully qualified for production use. This service is provided in collaboration with Google LLC. Google Cloud Platform support details can be found here.

Please review and accept the Terms of Service below to start or continue use of Cloud Dataprep.

You must accept the Trifacta Terms of Service.

☑ By checking this box, you agree to the Trifacta Terms of Service.



Welcome, student 77165863!

Example Flow - Sales Analytics Use Case Today at 8:14 PM

Example Flow - Event log data Today at 8:14 PM

The browser you are using (Edge 137.0.0.0) is not officially supported.

Your trial ends in 30 days. Subscribe now to continue with Dataprep!

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

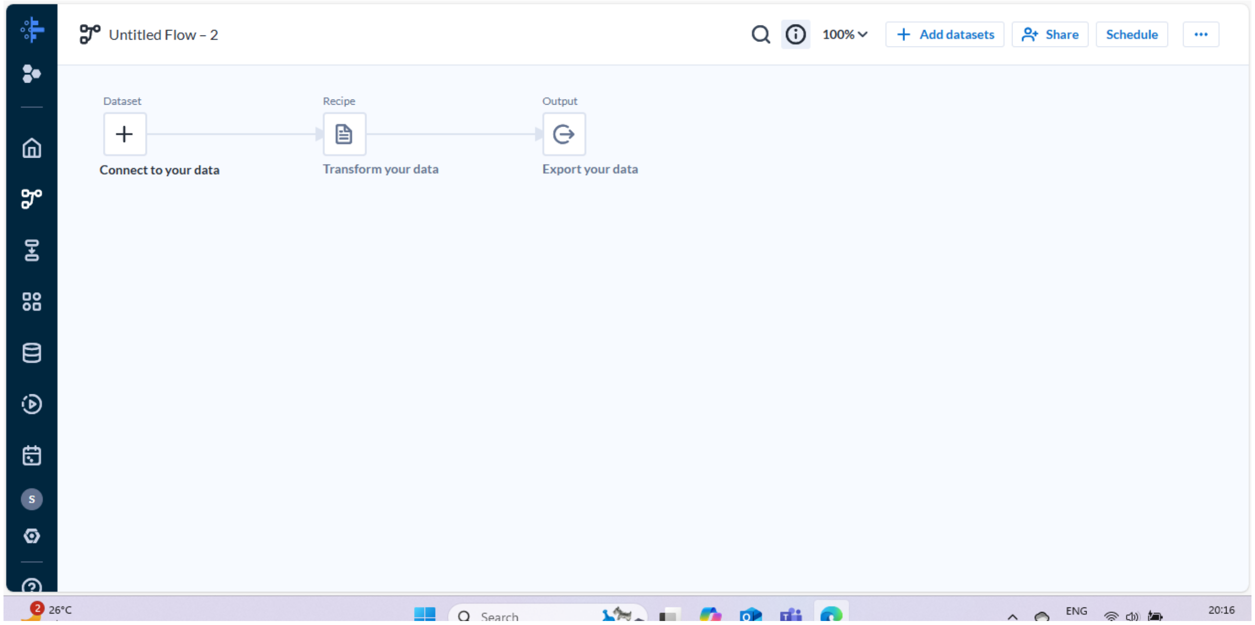
You may have to widen the browser window to see the **Go** and **Cancel** buttons.

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

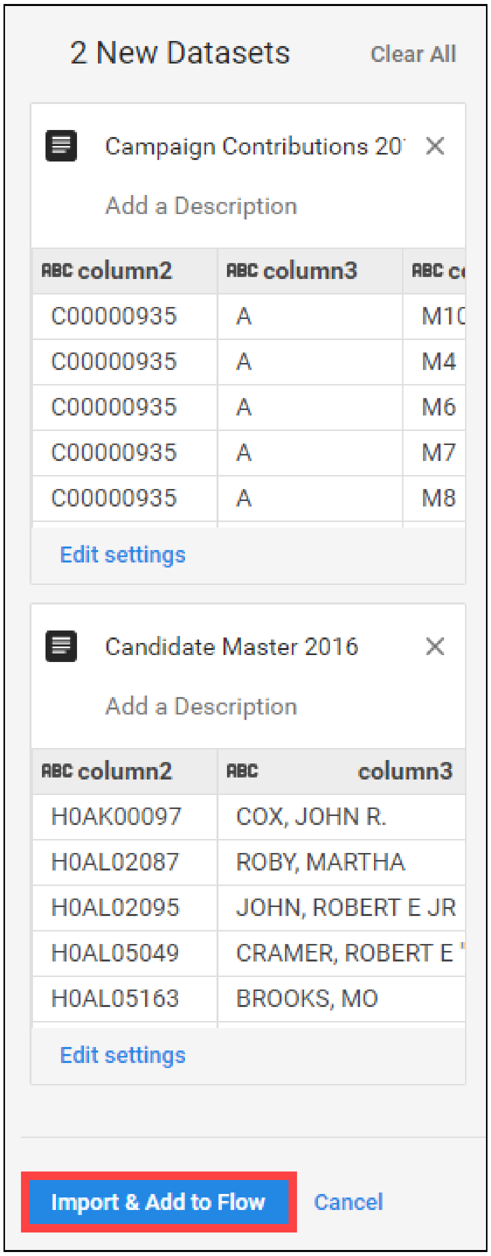
5. Click the **+** icon next to cn-2016.txt to create a dataset shown in the right pane. Click on the title in the dataset in the right pane and rename it "Candidate Master 2016".

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**.



Untitled Flow – 2
100%
Add datasets
Share
Schedule
···
Dataset
Recipe
Output
Connect to your data
Transform your data
Export your data
S
2
26°C
Search
ENG
20:16



Campaign Contributions 20

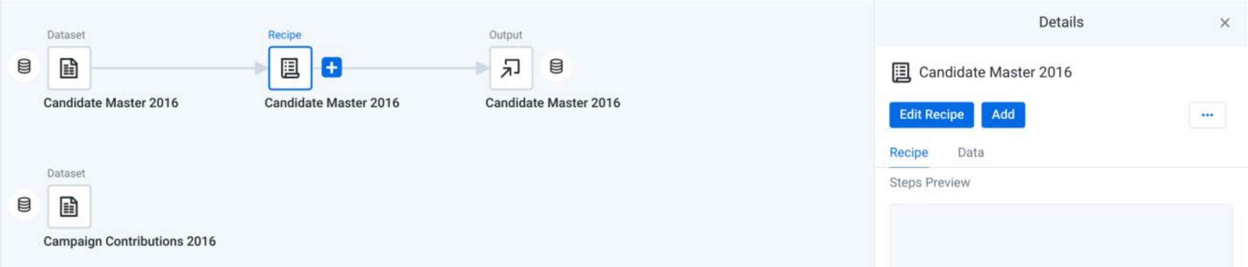
ABC column2	ABC column3	ABC c
C00000935	A	M1
C00000935	A	M4
C00000935	A	M6
C00000935	A	M7
C00000935	A	M8

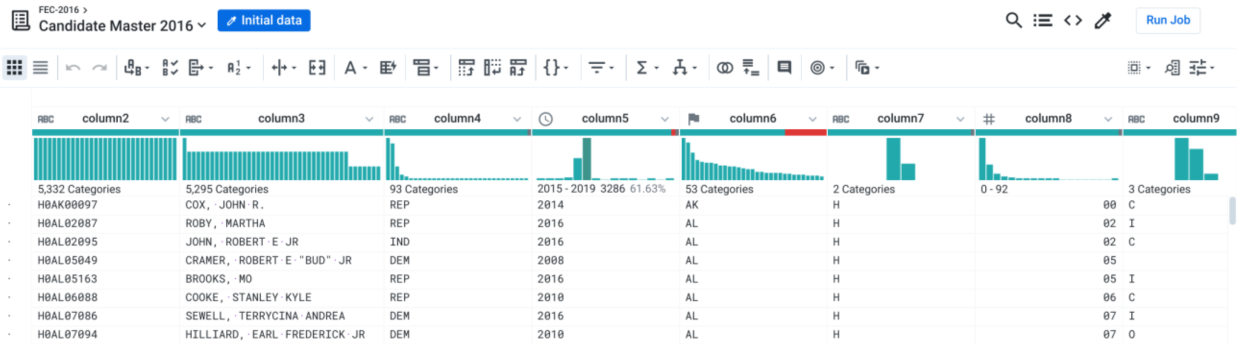
Candidate Master 2016

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

Task 5. Prep the candidate file

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

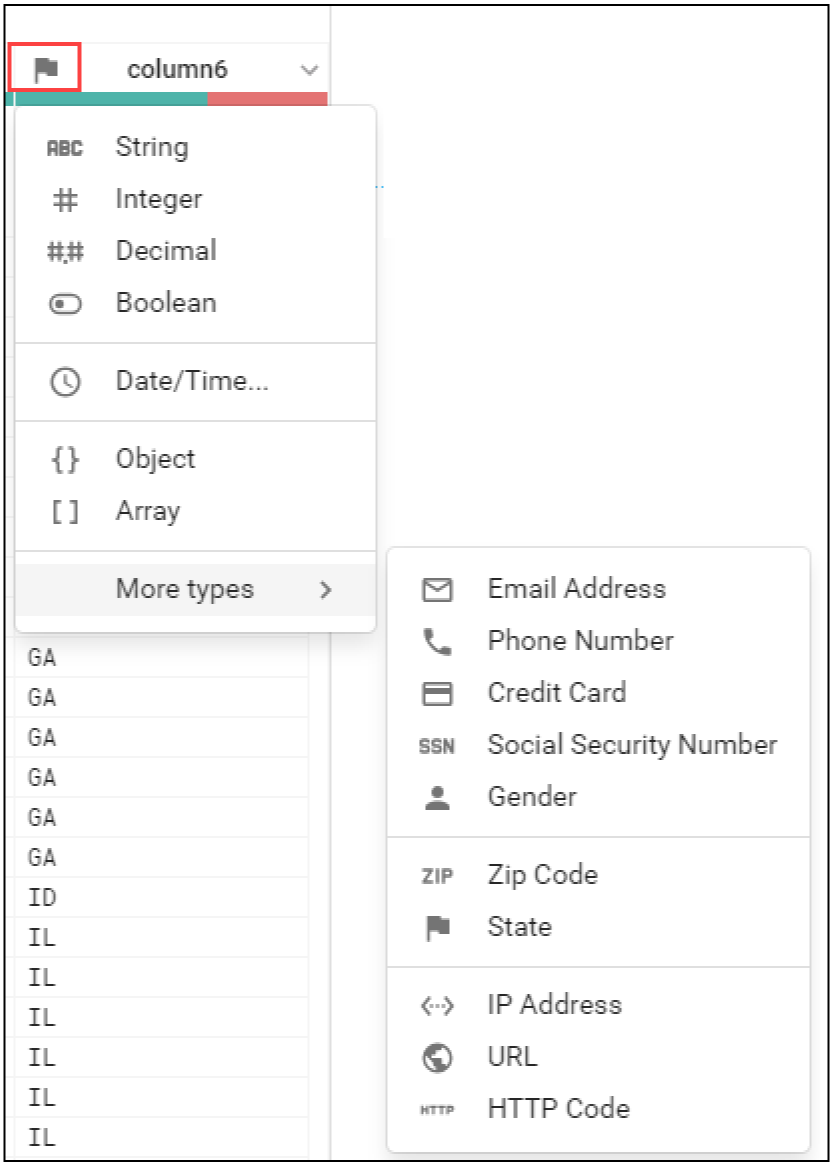




column2	column3	column4	column5	column6	column7	column8	column9
5,332 Categories	5,295 Categories	93 Categories	2015 - 2019 3286 61.63%	53 Categories	2 Categories	0 - 92	3 Categories
H0AK00097	COX, JOHN R.	REP	2014	AK	H	00	C
H0AL02087	ROBY, MARTHA	REP	2016	AL	H	02	I
H0AL02095	JOHN, ROBERT E JR	IND	2016	AL	H	02	C
H0AL05049	CRAMER, ROBERT E "BUD" JR	DEM	2008	AL	H	05	
H0AL05163	BROOKS, MO	REP	2016	AL	H	05	I
H0AL06088	COOKE, STANLEY KYLE	REP	2010	AL	H	06	C
H0AL07086	SEWELL, TERRYCINA ANDREA	DEM	2016	AL	H	07	I
H0AL07094	HILLIARD, EARL FREDERICK JR	DEM	2010	AL	H	07	O

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:

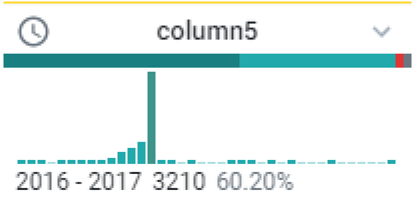


3. Notice also that when you click the name of the column, a **Details** panel opens on the right.

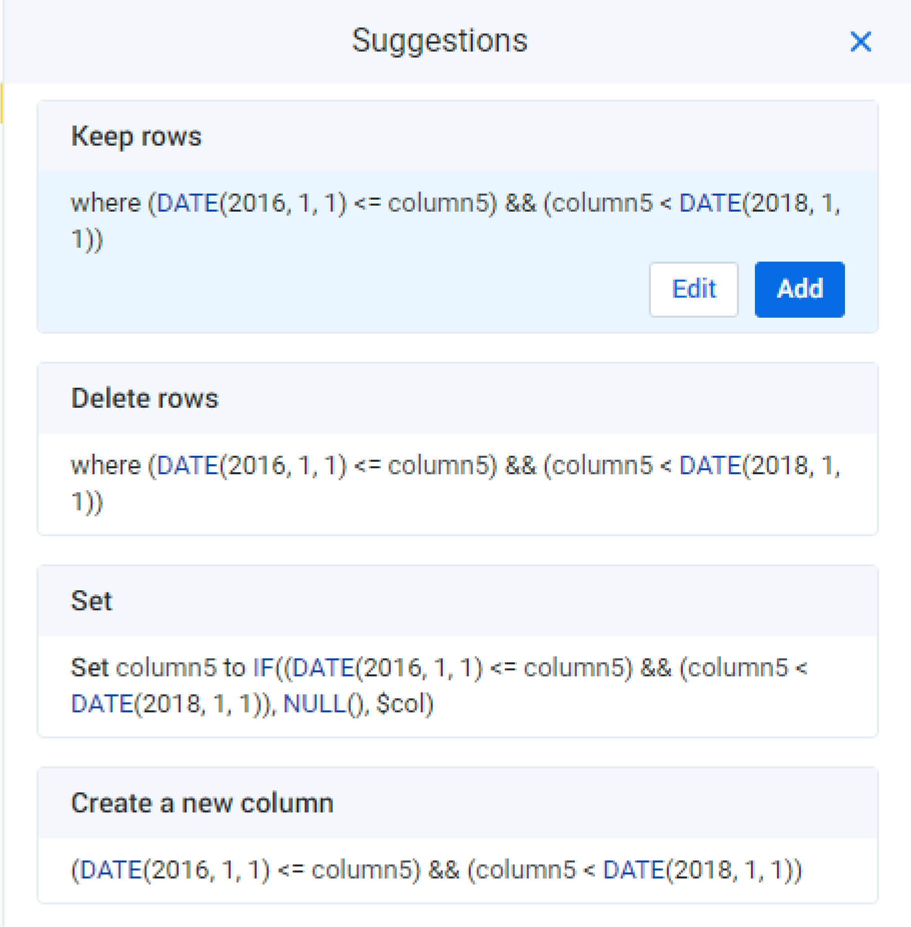
4. Click **X** in the top right of the Details panel to close the **Details** panel.

   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.



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.

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.

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.

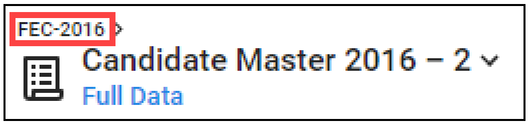
6. In the right Suggestions panel, click **Add** to accept the step to the recipe.

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.

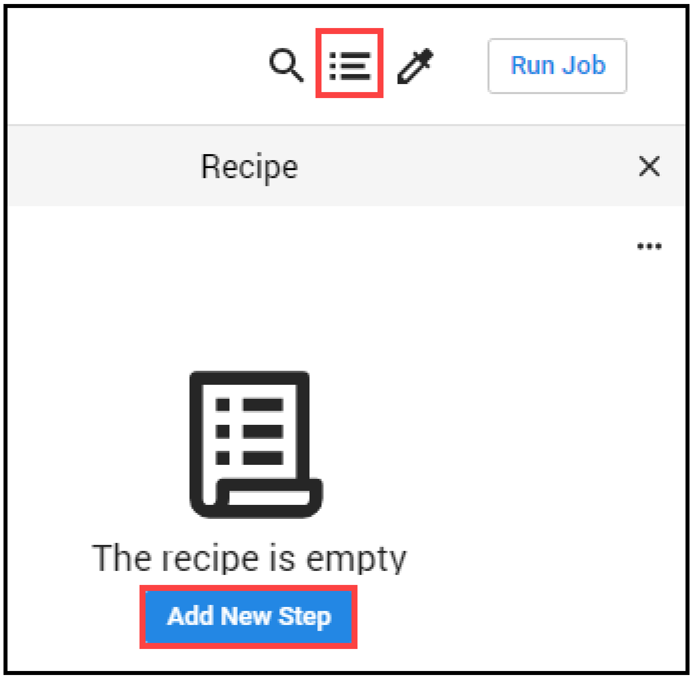
1. Click on **FEC-2016** (the dataset selector) at the top of the grid view page.



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

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

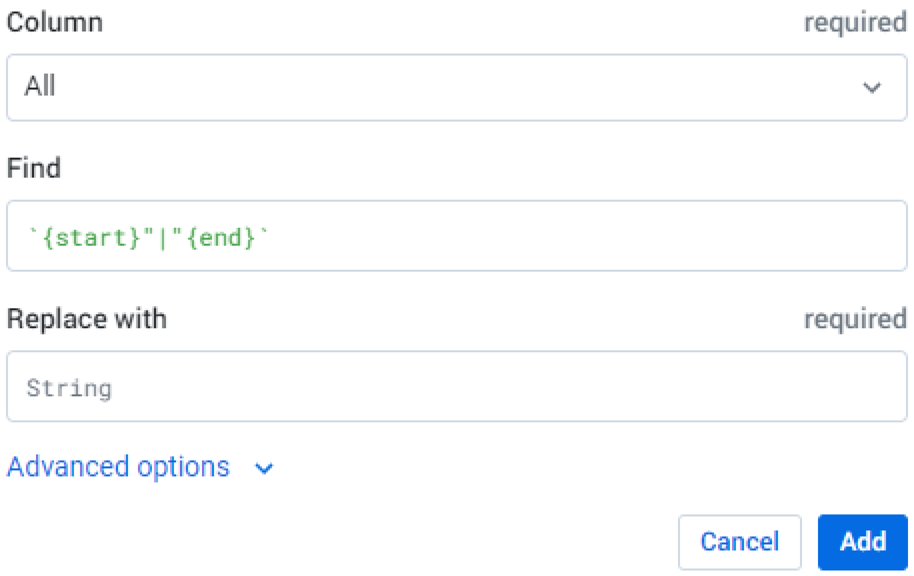
4. Click the **recipe** icon at the top right of the page, then click **Add New Step**



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
```



Column required

All

Find

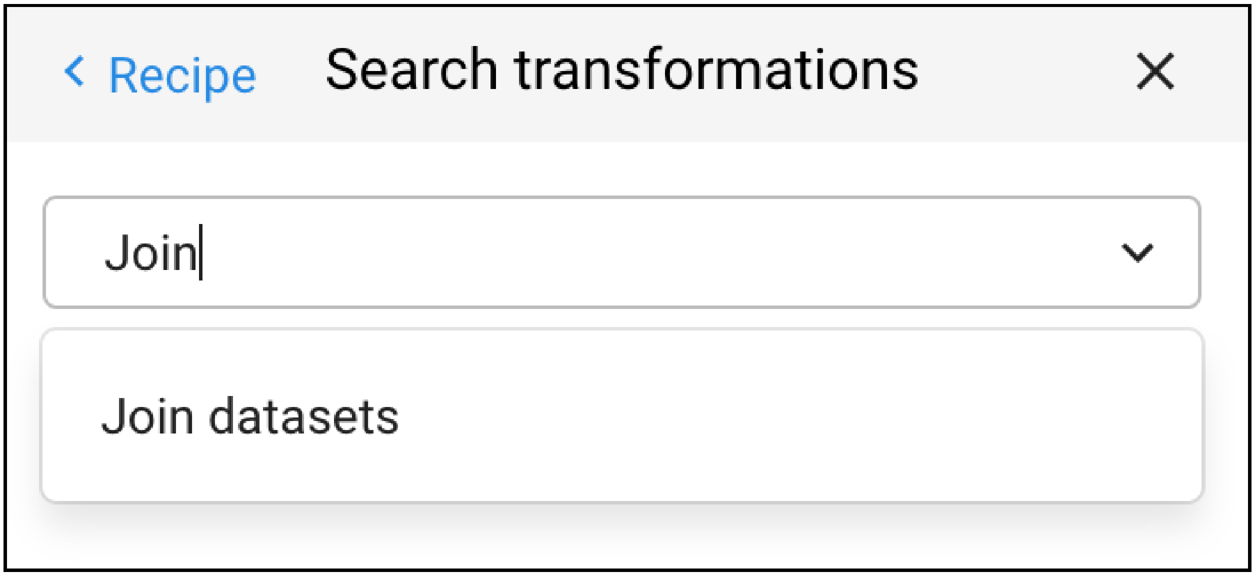
`{start}"|"{end}`

Replace with required

String

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

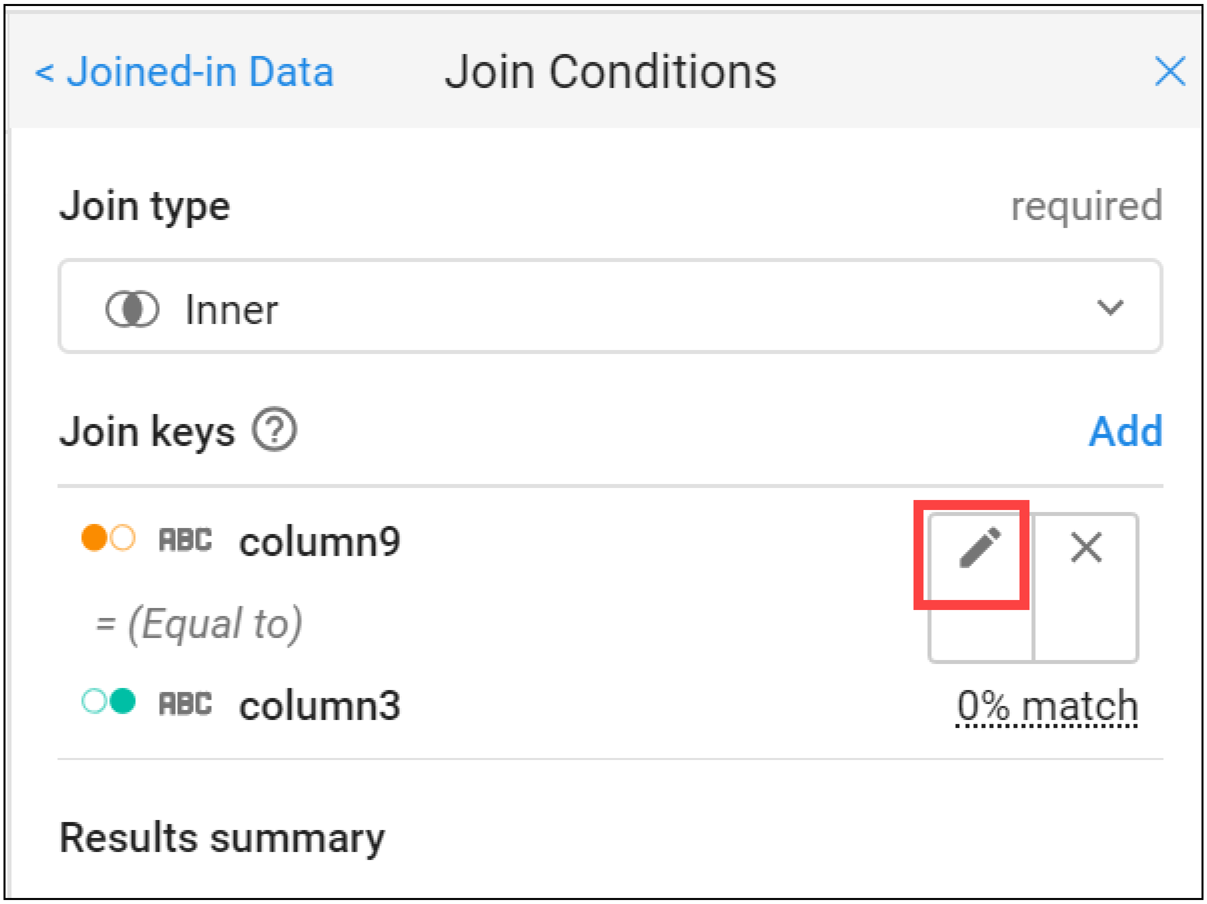
7. Add another new step to the recipe. Click **New Step**, then type "Join" in the Search box.



Join

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.



< Joined-in Data
Join Conditions

Join type: Inner (required)

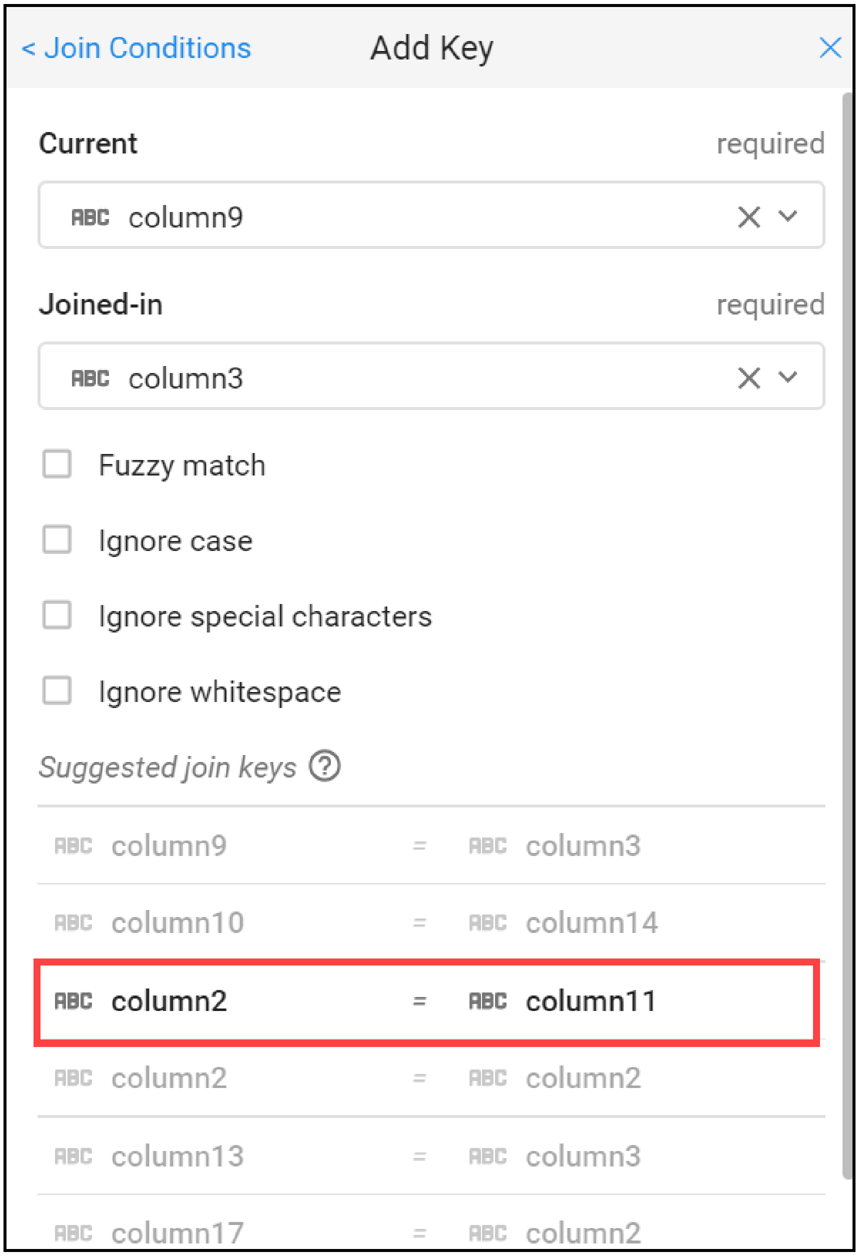
Join keys

ABC column9
= (Equal to)
ABC column3

0% match

Results summary

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

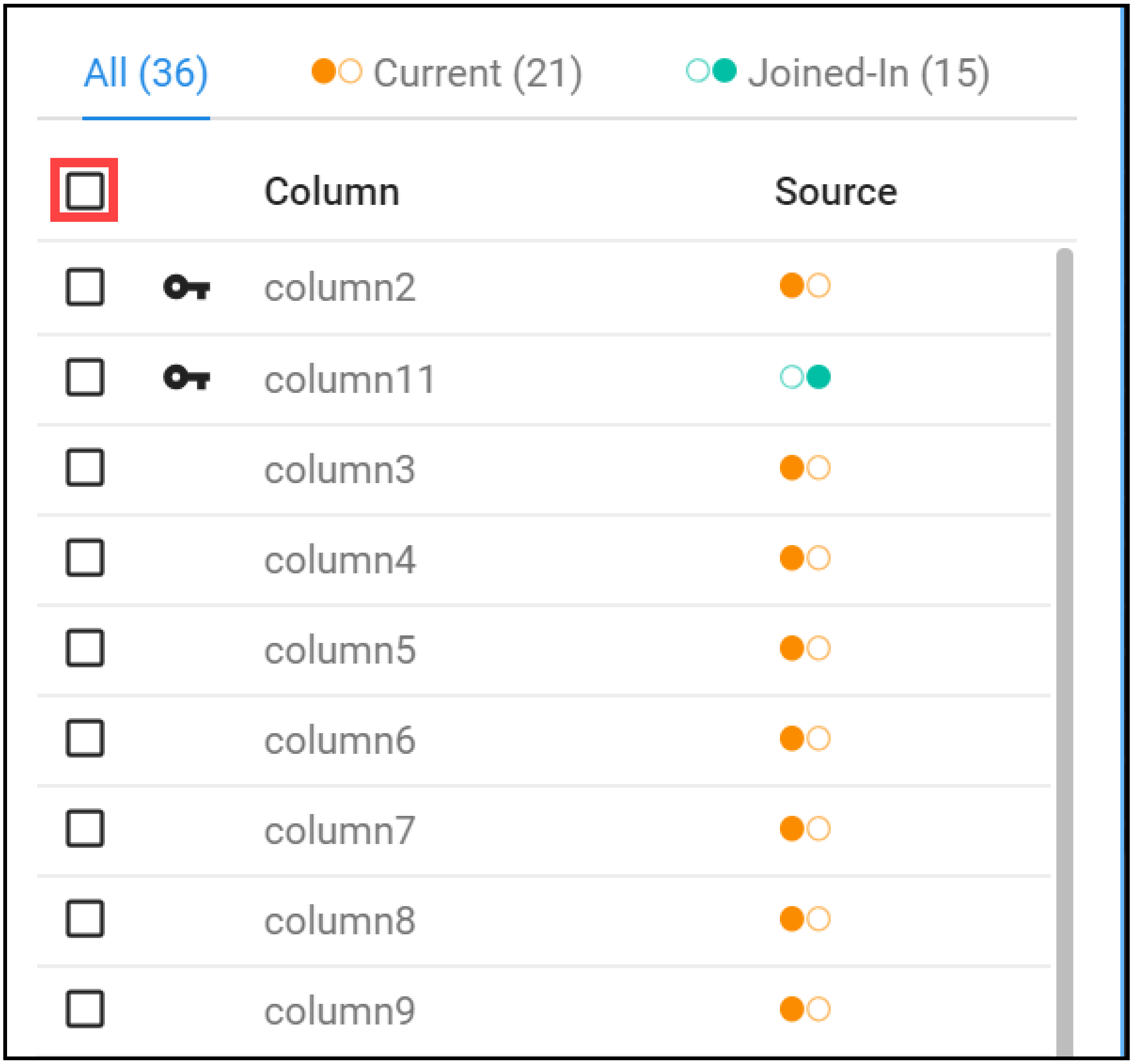
Suggested join keys ?

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)

□		Column	Source
□	🔑	column2	●○
□	🔑	column11	○●
□		column3	●○
□		column4	●○
□		column5	●○
□		column6	●○
□		column7	●○
□		column8	●○
□		column9	●○

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

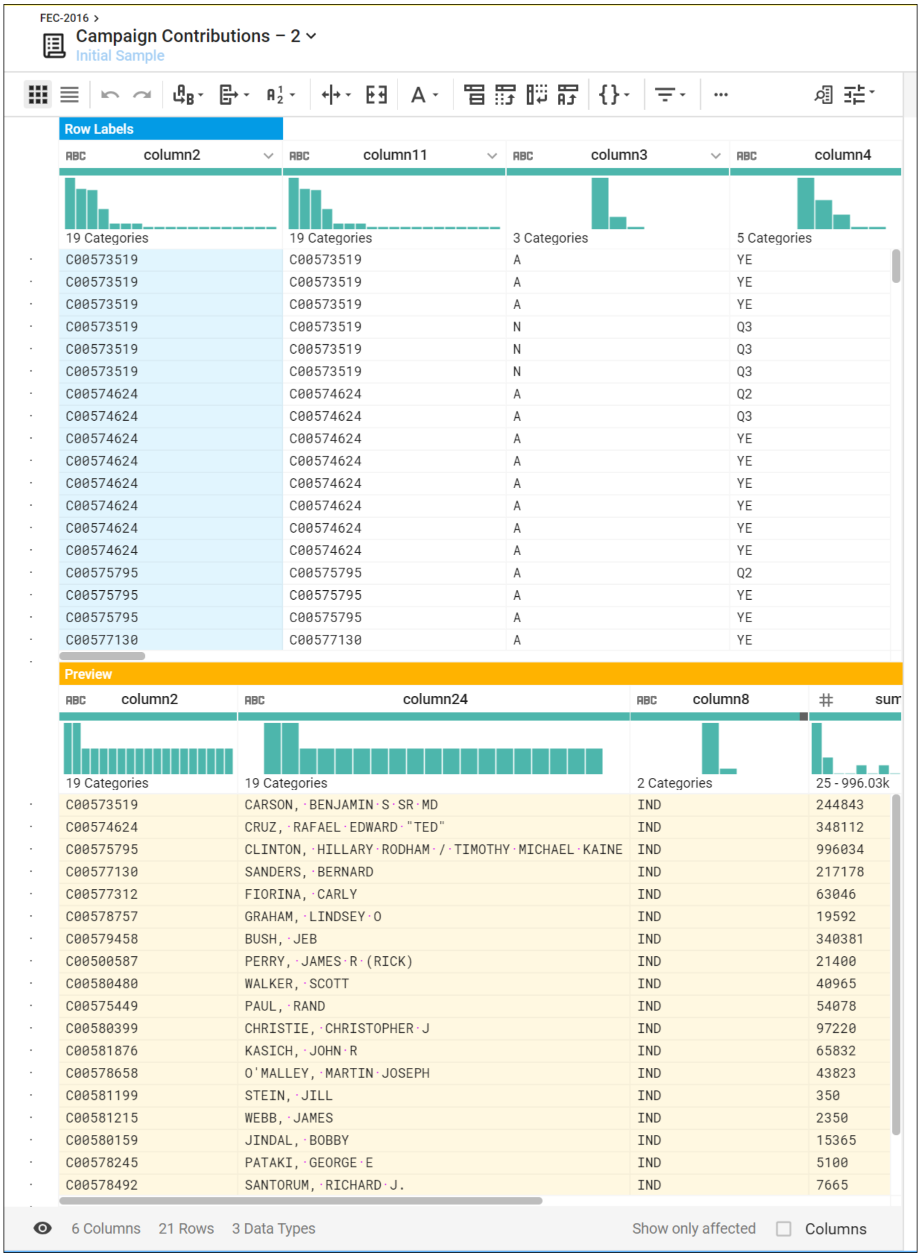
Task 7. Summary of data

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.


FEC-2016 >

Campaign Contributions – 2

Initial Sample

Row Labels

column2	column11	column3	column4
19 Categories	19 Categories	3 Categories	5 Categories
C00573519	C00573519	A	YE
C00573519	C00573519	A	YE
C00573519	C00573519	A	YE
C00573519	C00573519	N	Q3
C00573519	C00573519	N	Q3
C00573519	C00573519	N	Q3
C00574624	C00574624	A	Q2
C00574624	C00574624	A	Q3
C00574624	C00574624	A	YE
C00574624	C00574624	A	YE
C00574624	C00574624	A	YE
C00574624	C00574624	A	YE
C00574624	C00574624	A	YE
C00574624	C00574624	A	YE
C00575795	C00575795	A	Q2
C00575795	C00575795	A	YE
C00575795	C00575795	A	YE
C00577130	C00577130	A	YE

Preview

column2	column24	column8	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
C00577312	FIORINA, CARLY	IND	63046
C00578757	GRAHAM, LINDSEY O	IND	19592
C00579458	BUSH, JEB	IND	340381
C00500587	PERRY, JAMES R (RICK)	IND	21400
C00580480	WALKER, SCOTT	IND	40965
C00575449	PAUL, RAND	IND	54078
C00580399	CHRISTIE, CHRISTOPHER J	IND	97220
C00581876	KASICH, JOHN R	IND	65832
C00578658	O'MALLEY, MARTIN JOSEPH	IND	43823
C00581199	STEIN, JILL	IND	350
C00581215	WEBB, JAMES	IND	2350
C00580159	JINDAL, BOBBY	IND	15365
C00578245	PATAKI, GEORGE E	IND	5100
C00578492	SANTORUM, RICHARD J.	IND	7665

6 Columns 21 Rows 3 Data Types

Show only affected □ Columns

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:

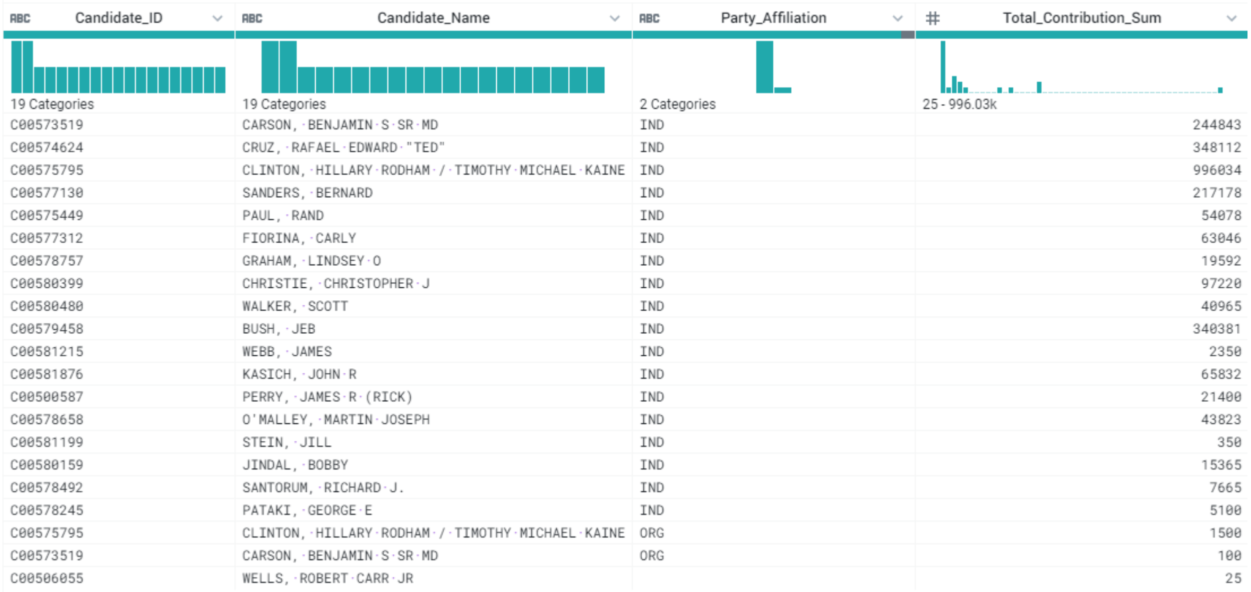
rename type: manual mapping: [column24,'Candidate_Name'], [column2,'Candidate_ID'],[column8,'Party_Affiliation'], [sum_column16,'Total_Contribution_Sum'], [average_column16,'Average_Contribution_Sum'], [countif,'Number_of_Contributions']

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**.



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