

Explore Data Flows in Oracle Analytics

Before You Begin

This lab shows you how to create a data flow with two datasets, add columns, transform data in the columns, and create new columns using expressions.

Background

In the tutorial, you create a data flow with a spreadsheet file containing sample school donation data (donation.xlsx) and a postal code statistics (zip_stats.xlsx) data file. In the data flow, you modify the donation dataset by filtering data in a column, and then select and add columns. From the two datasets, you create a curated dataset that you can use in analyses.

What Do You Need?

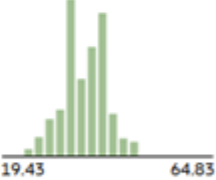
- Access to Oracle Analytics Cloud or Oracle Analytics Desktop
- Download the following source files:
 - donation.xlsx
 - zip_stats.xlsx

Create a Dataset

In this section, you create a dataset using the donation.xlsx file. When numerical data is loaded, it is treated as a measure. You learn how change the Treat as value for numerical columns that are attributes.




1. Sign in to Oracle Analytics.
2. On the Home page, click **Create**, and then click **Dataset**.
3. In Create Dataset, click **Drop data file here or click to browse**. In File Upload, select the donation.xlsx file, and then click **Open**.
4. In Create Dataset Table from donation.xlsx, click **OK**.

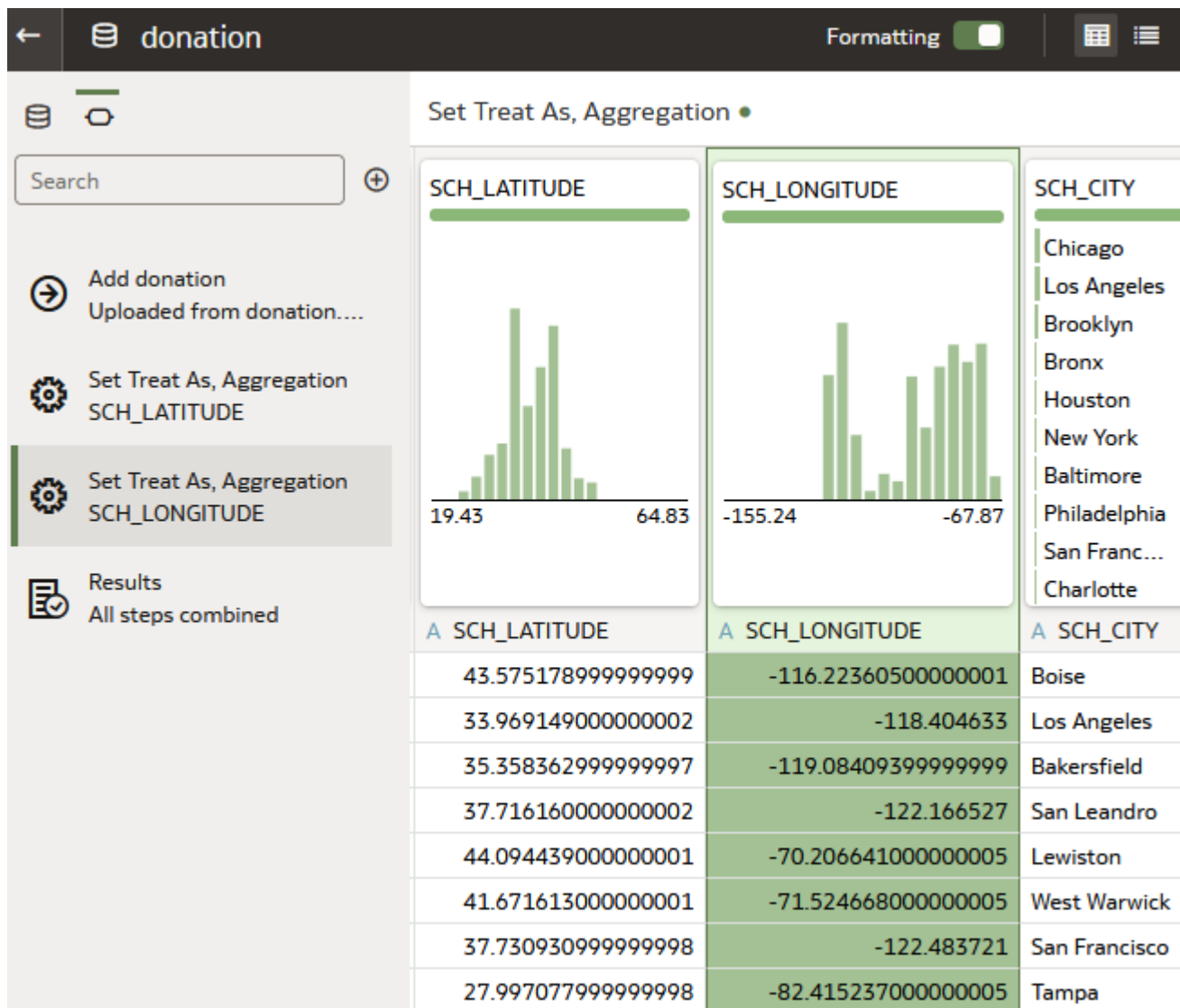
donation

PROJECTID	TEACHER_ACCTID	SCHOOL_ID	SCH_LATITUDE
This column contains 100% unique values.	This column contains 98.90% unique values.	5de075e3c3852d4fd286c2678da3e... 11d822fcc3ddddd7159fa072288ec8... 2b3e27e5b22025e72ce927e375cc8... 3264a9d682b91e4801f148eb6cae6... 49f2955657a66fd001967c97bc793... 06eaabb3df286b5f10d7fbb1648d5... 0c74736934d9fffe825609f25b53a... 12fa1d813e512f1988a13f7a305f3... 224dd526a94948eeef1d6b9223a6... 2b6b158f28b7922af77d46e3a905d...	
A PROJECTID	A TEACHER_ACCTID	A SCHOOL_ID	# SCH_LATITUDE
P26626	31a8d0415addc20f918faeb021bf76dd	58d2729b69d82986c6090d87c009ec57	43.5751790000000000
P208212	c0980f7cebada9f53aba72e017ab116b	72df35952cda0a12156a41282ab8566c	33.9691490000000000
P404052	359f8763a3d49fd2b45cf214345c429f	202d0a51024056ed479a8c27350d086e	35.3583630000000000
P234238	a9557f8f0add814720a978458a90f4ba	36485fe4828b4eaf5ec00cbd240f5817	37.7161600000000000
P288233	5ddd5b0dfee9164725c98bbe4899c1c7	af3462a5d2af955b2597439133a083bc	44.0944390000000000
P255227	9f009cab416972b59d1755c5efaeab55	d5ede2180d7f53f29bb1a8d80d4c655f	41.6716130000000000

5. Click the donation tab.

Join Diagram  **donation**

6. Scroll to the **SCH_LATITUDE** column, click **Measure** , and then click **Attribute**.
7. Scroll to the **SCH_LONGITUDE** column, click **Measure** , and then select **Attribute**.
8. Click **Save** . In Save Dataset As, enter donation in **Name**, and then click **Save**.

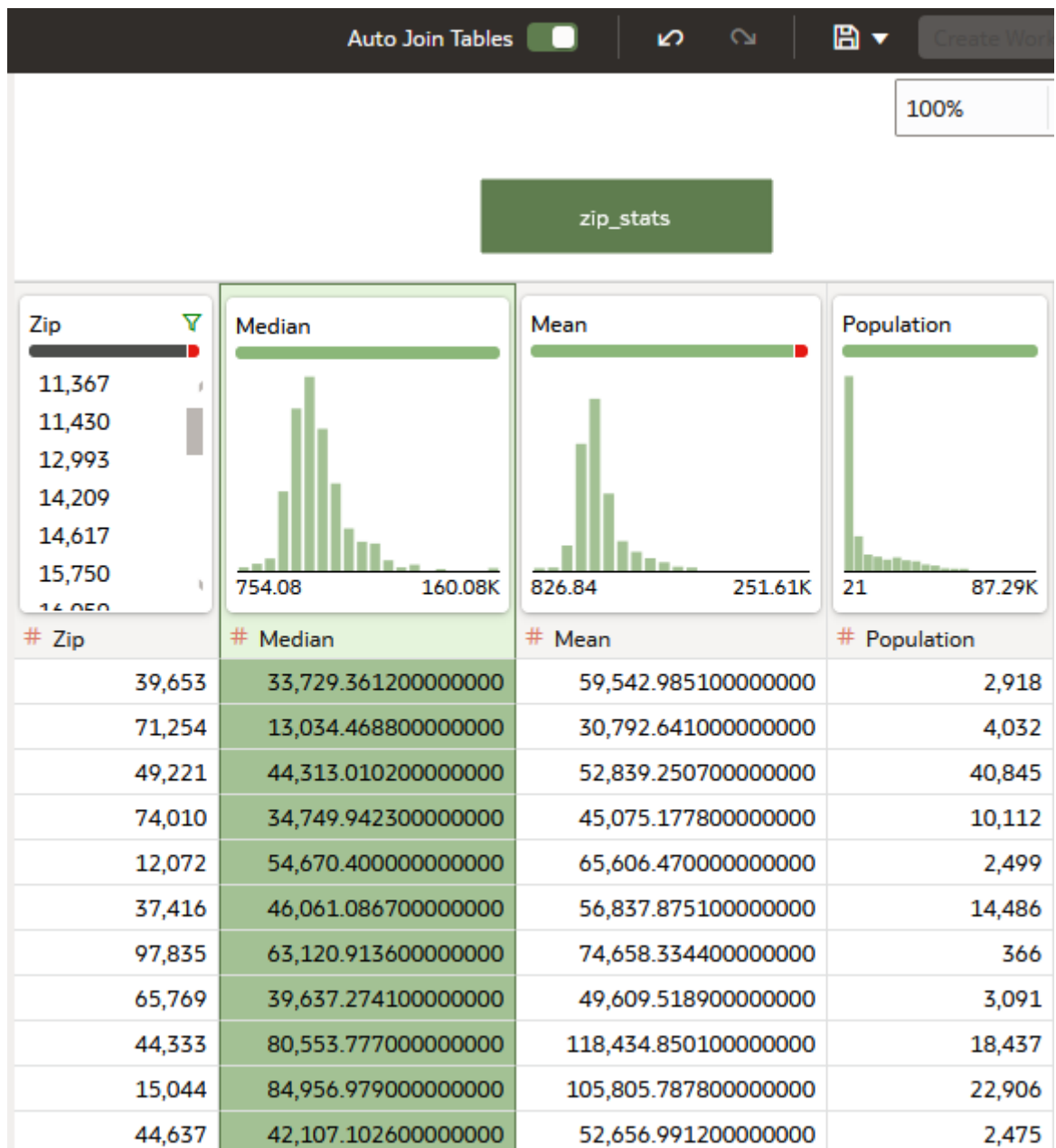


9. In the donation dataset page, click **Go back** .

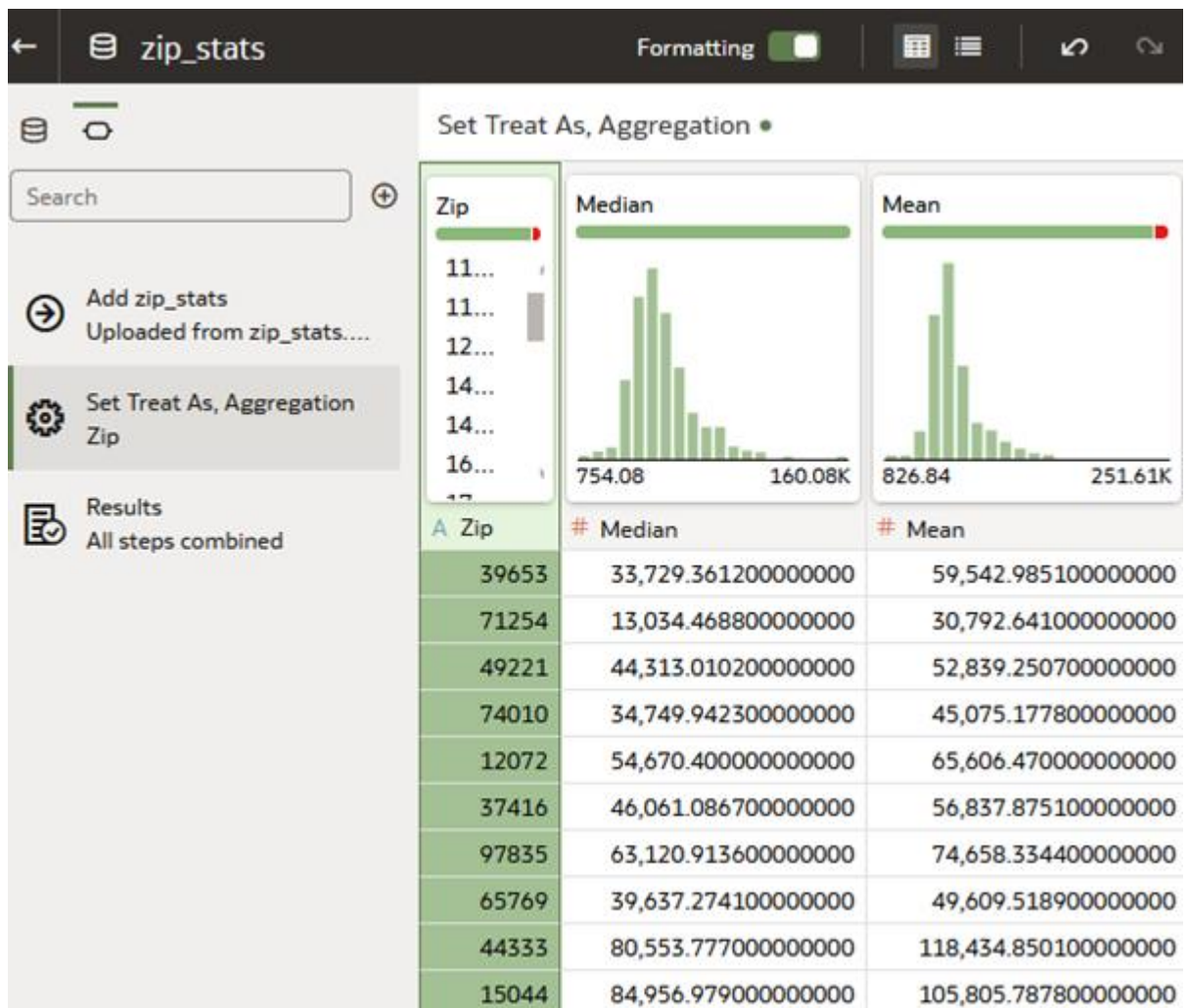
Add a Second Dataset

In this section, you create a dataset that provides demographics to add to the donation dataset.

1. On the Home page, click **Create**, and then click **Dataset**.
2. In Create Dataset, click **Drop data file here or click to browse**. In File Upload, select the zip_stats.xlsx file, and then click **Open**.
3. In Create Dataset Table from zip_stats.xlsx, click **OK**.





- Click the **zip_stats** tab. In zip_stats, select the **Zip** column, click **Measure** #, and then select **Attribute**.
- In the zip column, click **Options** ⋮, and then click **Convert to Text**.
- Click **Save** . In Save Dataset As, enter zip_stats in **Name**, and then click **Save**.

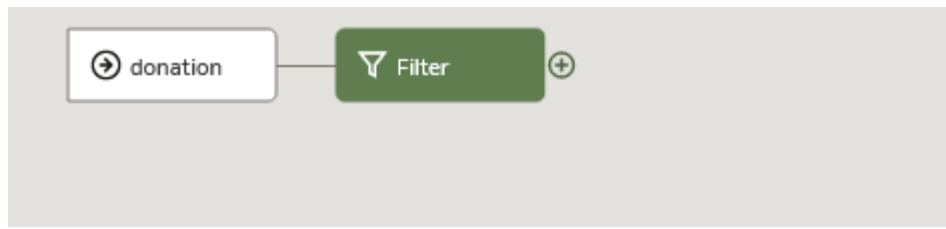


7. Click **Go back** .

Create a Data Flow

In this section, you create a data flow and implement a filter step to remove the donation records that don't have a value in the DATE_COMPLETED column.

1. On the Home page, click **Create**, and then click **Data Flow**.
2. In Add Dataset, select the **donation** dataset, and then click **Add**.
3. Click **Add a step**  on the donation node, and select **Filter**.
4. In Filter, click **Add Filter** , and then select **DATE_COMPLETED** from Available data.
5. In DATE_COMPLETED, under Date Range keep **Range** as the value. In the first field, enter 1/1/2010 as the start date. In the second field enter 12/31/2014 and the end date, and then click inside the dialog.



Filter

DATE_COMPLETED
1/1/2010 - 12/31/2014

DATE_COMPLETED

Date Range

Range 1/1/2010 - 12/31/2014

ab PROJECTID	ab TEACHER_ACCTID	ab SCHOOL_ID
P100000	c613d9f2d60fdb26e488e6a258ec325d	f5a61953a1bc3491
P100003	5b3fbcd26b5c906a6682595bd6045440	57a2826ad350c90

- Click **Save**. In Save Data Flow As, enter School Donations, and then click **OK**.

Add Columns

In this section, you create columns by defining expressions.

- From Data Flow Steps, drag the **Add Columns** to **Add a step** on the Filter node. In **Name**, enter SCH_STATE. In the expression field, enter the following:
Substring(SCH_STATEZIP from 1 for 2)
- Click **Validate**, and then click **Apply**.
- In Add Columns, click **Column** . In **Name**, enter SCH_ZIP. In the expression field, enter the following:
cast(Substring(SCH_STATEZIP from 4) as int)
- Click **Validate**, and then click **Apply**.
- In Add Columns, click **Column** . In **Name**, enter YR_COMPLETED. In the expression field, enter the following:
Year(DATE_COMPLETED)
- Click **Validate**, and then click **Apply**.
- Click **Toggle auto-refresh Data Preview** , and then scroll to view the new columns.
- Select the **SCH_ZIP** column, click **Options** , and then select **Convert to Text**. Select the **YR_COMPLETED** column, click **Options** , and then select **Convert to Text**.

The Transform column data flow steps are automatically added to the data flow.

donation → Filter → Add Columns → Transform Column → Transform Column

Transform Column

Transform **YR_COMPLETED**

Name: f(x)

Search Q

► Operators

IO...	ab FUNDING_ST...	ab DATE_POSTED	DATE_COMPL...	ab SCH_STATE	ab SCH_ZIP	ab YR_COMPLET...
	completed	10.Feb.2011	03/20/2011	NC	28212	2011
	expired	10.Feb.2011	02/28/2011	CT	6035	2011
	completed	10.Feb.2011	03/17/2011	TX	77093	2011
	completed	11.Feb.2011	06/22/2011	TN	38116	2011
	completed	11.Feb.2011	03/24/2011	TX	78521	2011
	expired	23.Feb.2010	03/09/2010	MI	48091	2010
	expired	11.Feb.2011	03/09/2011	FL	33584	2011
	expired	11.Feb.2011	04/14/2011	CA	95076	2011
	expired	11.Feb.2011	03/10/2011	GA	31406	2011

Add a Dataset to the Data Flow

In this section, you add the zip_stats spreadsheet to the data flow.

1. In Data Flow Steps, double-click **Add Data**. In Add Dataset, click **zip_stats**, and then click **Add**.
2. In the data flow, click the **Join** step. From Input 1 list, select **All Rows**. Under Match Columns, select **SCH_ZIP** for the Input 1 value.

The Input 2 value contains the Zip column from the zip_stats dataset.

donation → Filter → Add Columns → Transform Column → Transform Column → **Join**

zip_stats

Join

Keep rows

1 Input 1:

2 Input 2:

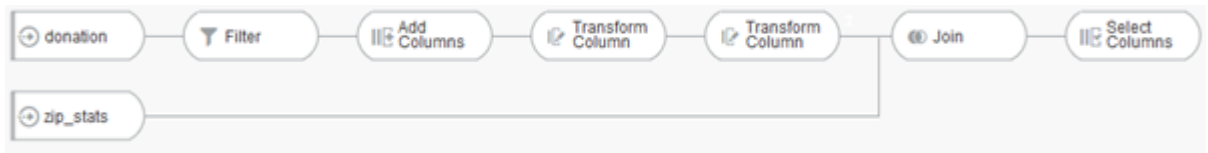
Match columns

Input 1	Operator	Input 2
SCH_ZIP	=	Zip



Select Columns

In this section, you select columns to use from the joined datasets.

1. From Data Flow Steps, drag **Select Columns** to **Add a step**  on the Join node.



Under **Select Columns**, click **Remove All**. Hold down the **Ctrl** key, and select the following columns, and then click **Add Selected**:

- **PROJECTID**
 - **SCHOOL_ID**
 - **PRIMARY_FOCUS_SUBJECT**
 - **RESOURCE_TYPE**
 - **POVERTY_LEVEL**
 - **GRADE_LEVEL**
 - **STUDENTS_REACHED**
 - **TOTAL_DONATIONS**
 - **NUM_DONORS**
 - **SCH_STATE**
 - **SCH_ZIP**
 - **YR COMPLETED**
 - **Median**
 - **Population**
2. From Data Flow Steps, drag **Aggregate** to **Add a step**  on the **Select Columns** node.
 3. Under **Group by**, click **Remove**  next to the **PROJECTID** row. Click **Add Aggregate**, click in the new field, and then select **PROJECTID** from Available Data.
 4. Under **Aggregate**, from the **Function** list, select **Average** in the following rows:
 - **STUDENTS_REACHED**
 - **TOTAL_DONATIONS**
 - **NUM_DONORS**
 - **Median**
 - **Population**
 5. Under **Aggregate**, click **Add Aggregate**, select **PROJECTID**. From the **Function** list, select **Count**, and then enter Number of Projects in **New column name**. In the **Median** row, enter Income in **New column name**.

Aggregate

Aggregate	Function	New column name
STUDENTS_REACHED	Average	STUDENTS_REACHED Average
TOTAL_DONATIONS	Average	TOTAL_DONATIONS Average
NUM_DONORS	Average	NUM_DONORS Average
Median	Average	Income
Population	Average	Population Average
PROJECTID	Count	Number of Projects

+ Add Aggregate

Group by

- SCHOOL_ID
- PRIMARY_FOCUS_SUBJECT
- RESOURCE_TYPE
- POVERTY_LEVEL
- GRADE_LEVEL
- SCH_STATE
- SCH_ZIP
- YR_COMPLETED

+ Add Group

Define Aggregation

In this section, you use the Aggregate step to set the functions to use for some columns in the dataset.

1. From Data Flow Steps, drag **Add Columns** to **Add a step** on the Aggregate node.
2. In **Name**, enter Donation by Population.
3. To represent the average donation amount by the average population in the zip code, in the expression field, enter the following:
TOTAL_DONATIONS Average/Population Average
4. Click **Validate**. Click **Apply**.
5. In Add Columns, click **Column** . In **Name**, enter Average School Donation by Income.
6. In the expression field, enter the following:
 $\text{avg}(\text{TOTAL_DONATIONS Average by SCHOOL_ID})/\text{Income}$
7. Click **Validate**. Click **Apply**.

Add Columns

Column Name: *f(x)*

Donation by Population

Average School Donator

`avg(TOTAL_DONATIONS Average by SCHOOL_ID)/Income`


Search

- Operators
- Aggregate
- String
- Math




ab SCH_STATE	ab SCH_ZIP	ab YR_COMPLET...	99 Income	99 Population Av...	99 Donation by Popu...	99 Average School D...
TX	77384	2014	87,349.615900000000	11,857	0.0000000000000000	0.0039413698538901
AZ	85706	2013	32,439.039600000000	55,209	0.0085951565867884	0.0106130498036429
AZ	85706	2014	32,439.039600000000	55,209	0.0000000000000000	0.0106130498036429
AZ	85706	2014	32,439.039600000000	55,209	0.0084451810393233	0.0106130498036429
AZ	85706	2012	32,439.039600000000	55,209	0.0032170479450814	0.0106130498036429
AZ	85706	2012	32,439.039600000000	55,209	0.0092944990852941	0.0106130498036429
AZ	85706	2013	32,439.039600000000	55,209	0.0140993316307124	0.0106130498036429
SC	29045	2012	57,219.653100000000	21,871	0.0224123268254767	0.0085666370458999
CA	95122	2014	55,999.996500000000	56,545	0.0133819082146963	0.0152960128607627
CA	95122	2011	55,999.996500000000	56,545	0.0161814484039261	0.0152960128607627

Save and Run the Data Flow

In this section, you name the dataset that is output as a result of running the data flow, and examine the columns in the new dataset. You use the Donations by School dataset in the next tutorial.

1. From Data Flow Steps, drag **Save Dataset** to the last Add a step  node on the data flow.
2. In Save Dataset, enter Donations by School. From Save data to, select **Dataset Storage** to save the data in Oracle Analytics.
3. Click **Save**.



4. Click **Run Data Flow** .
5. After the data flow run completes successfully, click **Go back** .
6. On the Home page, select the **Donations by School** dataset, click the **Actions Menu** , and then select **Inspect**.
7. In Donations by School Dataset, click **Data Elements** to view the columns in the dataset.



Donations by School

Dataset

Save

General	Name	Data Type	Treat As	Aggregation
Data Elements	SCHOOL_ID	Text	Attribute	None
Search	PRIMARY_FOCUS_SUBJECT	Text	Attribute	None
Access	RESOURCE_TYPE	Text	Attribute	None
	POVERTY_LEVEL	Text	Attribute	None
	GRADE_LEVEL	Text	Attribute	None
	SCH_STATE	Text	Attribute	None
	SCH_ZIP	Text	Attribute	None
	YR_COMPLETED	Text	Attribute	None
	STUDENTS_REACHED Average	Number	Measure	Average
	TOTAL_DONATIONS Average	Number	Measure	Average
	NUM_DONORS Average	Number	Measure	Average
	Income	Number	Measure	Average
	Population Average	Number	Measure	Average
	Number of Projects	Number	Measure	Sum
	Donation by Population	Number	Measure	Sum
	Average School Donation by Income	Number	Measure	Sum