

# Create a Random Sample Dataset and Train a Prediction Model

## Before you Begin

This lab shows you how to create a random dataset, train a predictive model, create a live scenario, and use the datasets and scenario in visualizations.

## Background

In Oracle Analytics, predictive models use several embedded machine learning algorithms to mine your datasets, predict a target values, or identify classes of records.

Oracle's machine learning functionality is for advanced data analysts who have an idea of what they're looking for in their data, are familiar with the practice of predictive analytics, and understand the differences between algorithms.

This is the first lab in *Train and Apply Predictive Models* in Oracle Analytics. Complete the labs in the order listed.


- Create a Random Sample Dataset and Train a Prediction Model
- Inspect and Modify the Prediction Model
- Apply a Predictive Model

## What Do You Need?

- Access to Oracle Analytics Cloud or Oracle Analytics Desktop  
When using Oracle Analytics Desktop, you must install machine learning (DVML) to use Diagnostics Analytics (Explain), Machine Learning Studio, or advanced analytics.
- Download donation.xlsx to your computer

## Create a Dataset

In this section, you create a Dataset using the donation file. When numerical data is loaded, it's treated as a measure. You, also, learn how to correct 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**, select the **donation.xlsx** file, and then click **Open**.
4. In Create Dataset Table from donation.xlsx, click **OK**.
5. Click **Save** . In Save Dataset As, enter donation in **Name**, and then click **OK**.

# Prepare the Dataset

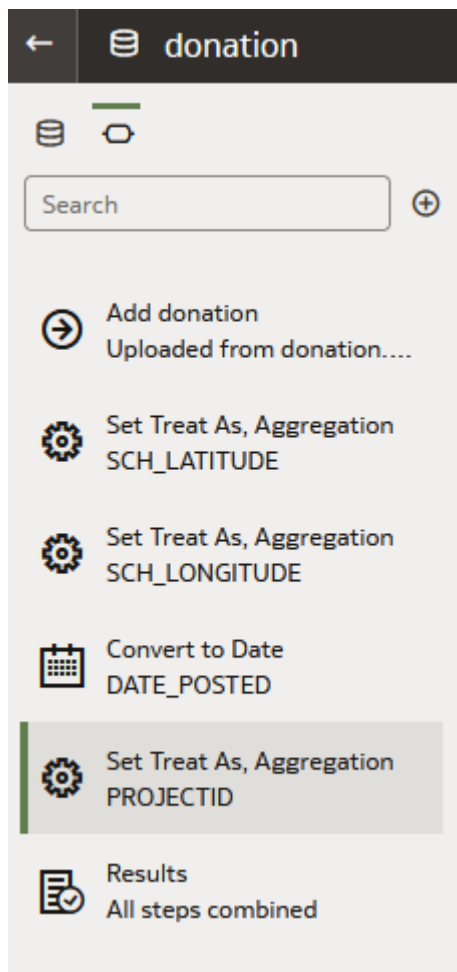
1. Click the **donation** tab.
2. Select **PROJECTID** column. In PROJECTID Properties, click **None** in the Aggregation row, and then select **Count**.

The screenshot shows the Power BI interface for the 'donation' dataset. The left sidebar contains a search bar, a 'Set Treat As, Aggregation' button for 'PROJECTID', and a 'Results' section showing 'All steps combined'. The main area displays the 'PROJECTID' column properties. The 'Treat As' row is set to 'Measure', the 'Data Type' is 'Text', and the 'Aggregation' is 'Count'. The 'Set Treat As, Aggregation' panel shows 'PROJECTID' with a message: 'This column contains 100% unique values.' and 'TEACHER\_ACCTID' with a message: 'This column contains 98.90% unique values.'

Name	PROJECTID	# PROJECTID	A TEACHER_ACCTID
Treat As	Measure	P26626	31a8d0415addc20f918faeb021bf76dd
Data Type	Text	P208212	c0980f7cebada9f53aba72e017ab116b
Aggregation	Count	P404052	359f8763a3d49fd2b45cf214345c429f
		P234238	a9557f8f0add814720a978458a90f4ba
		P288233	5ddd5b0dfee9164725c98bbe4899c1c7
		P255227	9f009cab416972b59d1755c5efaeab55

3. Scroll to the **SCH\_LATITUDE** column. In **SCH\_LATITUDE** properties, click **Measure** in the Treat As row, and then click **Attribute**. In the Data Type row, click **Number**, and then select **Text**.
4. Select the **SCH\_LONGITUDE** column. In **SCH\_LONGITUDE** properties, click **Measure** in the Treat As row, and then click **Measure**, and then select **Attribute**. In the Data Type row, click **Number**, and then select **Text**.
5. In the donation dataset page, click **Toggle Quality Insights** .
6. Right-click the **DATE\_POSTED** column, and then select **Convert to Date**. In Convert to Date, select **Custom** from the Source Format list, and then enter dd.MMM.yyyy as the date format.
7. Click **Save**.

The Preparation Script panel lists the changes that you've applied to the dataset.

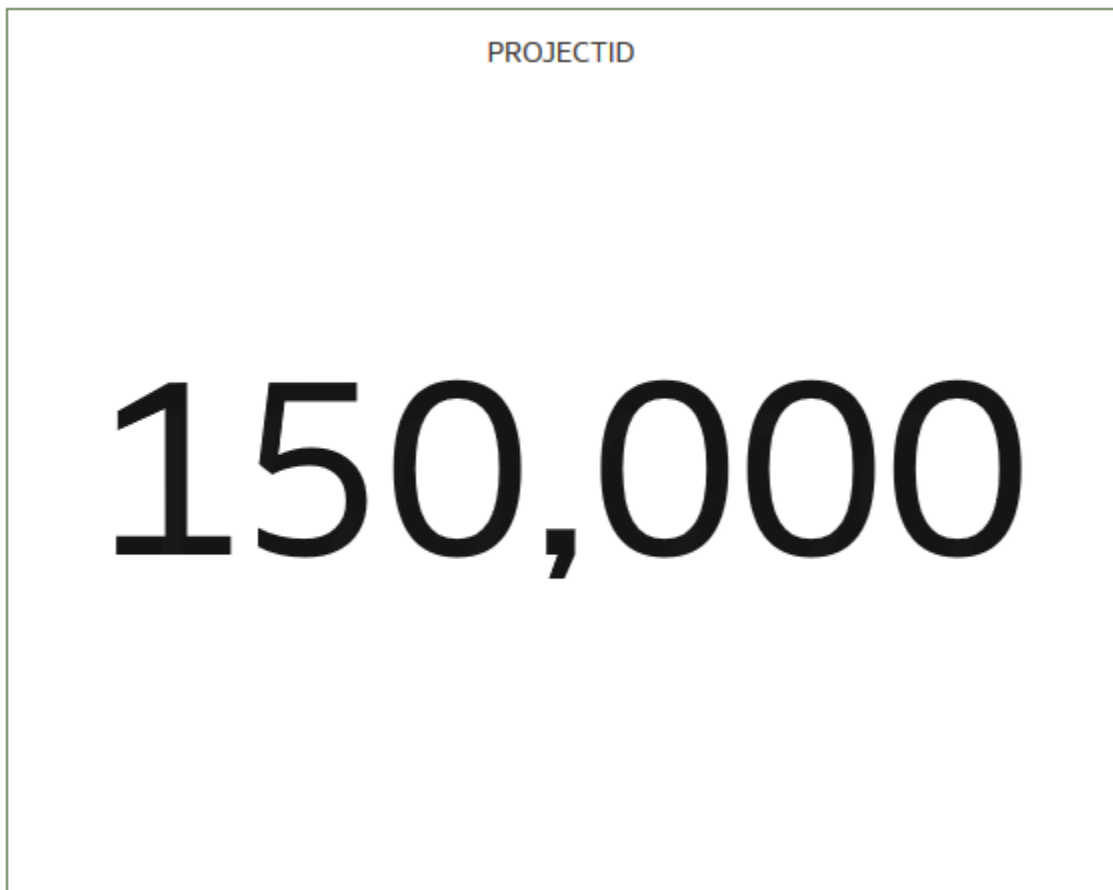


## Visualize the Data

In this section, you create visualizations with the donation dataset as a baseline to compare with the workbook that uses a random set of the donation data.

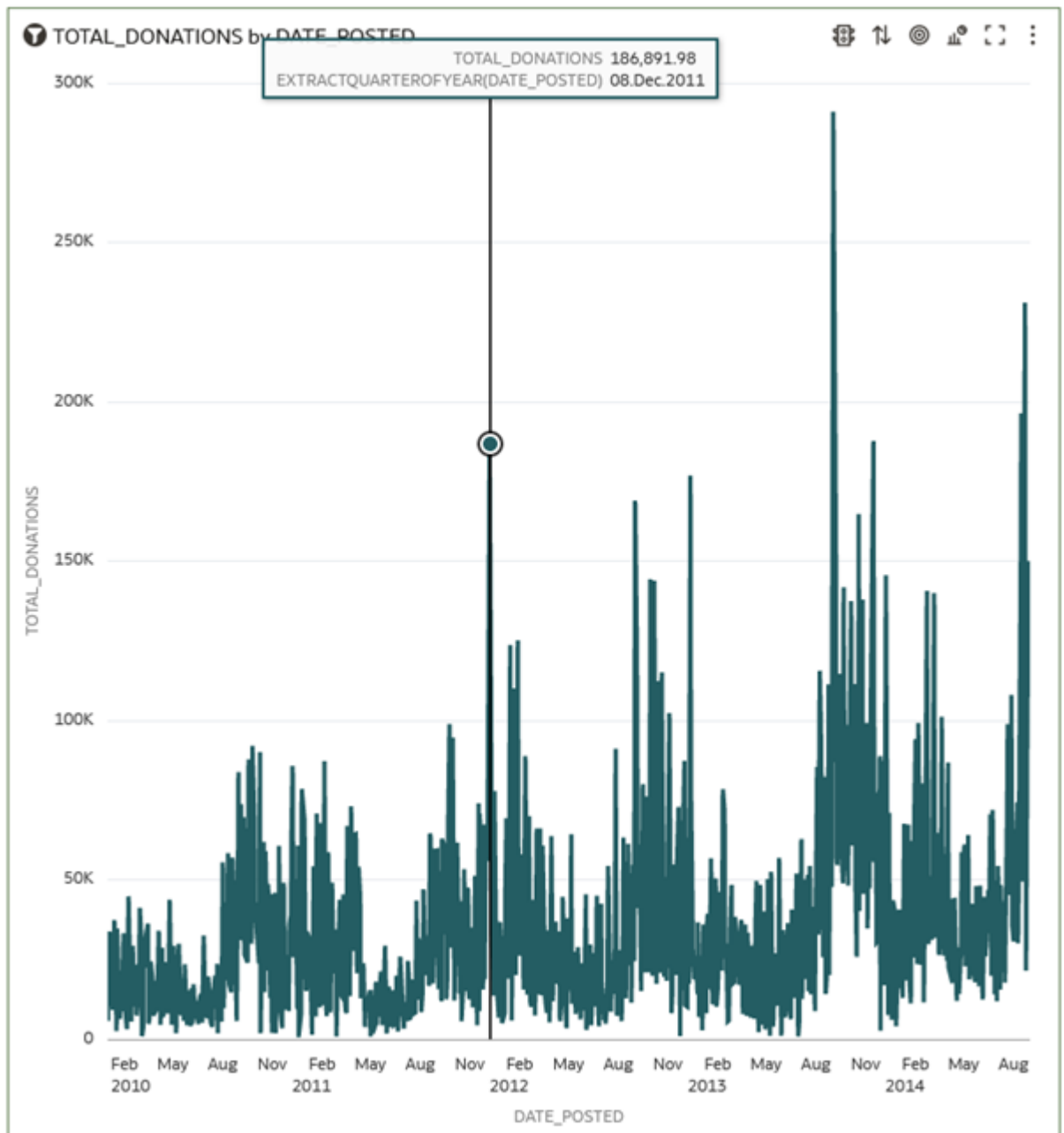
1. Click **Create Workbook**. Close the Auto Insights panel.
2. Drag **PROJECTID** from the Data panel, and then drop it on the visualization canvas.

The tile visualization shows the number of projects in the dataset.



3. In the Data panel, hold down the **Ctrl** key, select **TOTAL\_DONATIONS** and **DATE\_POSTED**. Right-click and then select **Create the Best Visualization**.

A line visualization is created with TOTAL\_DONATIONS on the Y-axis and DATE\_POSTED on the X-axis.



4. In the Data panel, select **TOTAL\_DONATIONS**, drag it to the canvas, and then drop **TOTAL\_DONATIONS** when a thick green line appears under the PROJECTID visualization.
5. Select the TOTAL\_DONATIONS by DATE\_POSTED line visualization, right-click **DATE\_POSTED** in Category (X-Axis), select **Show By**, and then select **Month**.

In the line visualization, you can begin to see patterns in donations data.

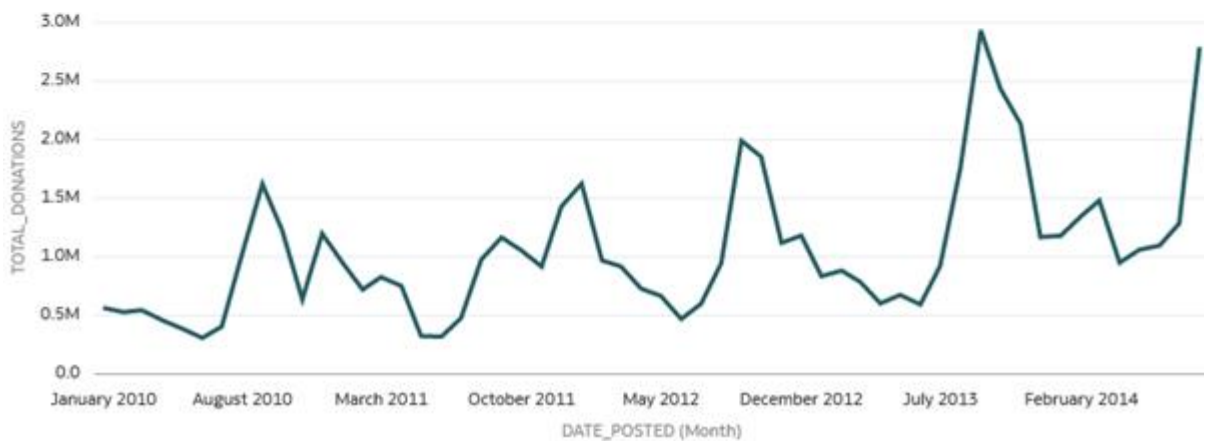
TOTAL\_DONATIONS



58,500,000.00

PROJECTID


150,000

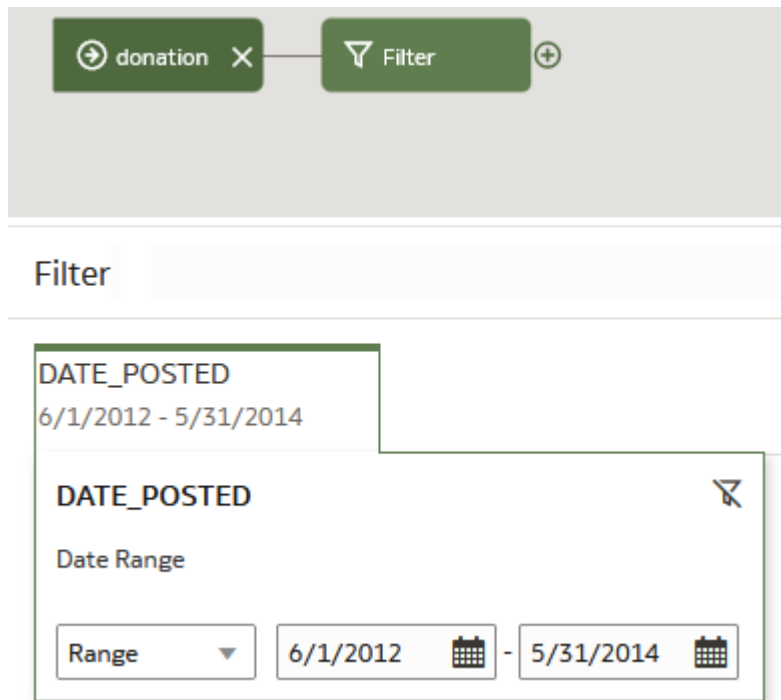
TOTAL\_DONATIONS by DATE\_POSTED (Month)



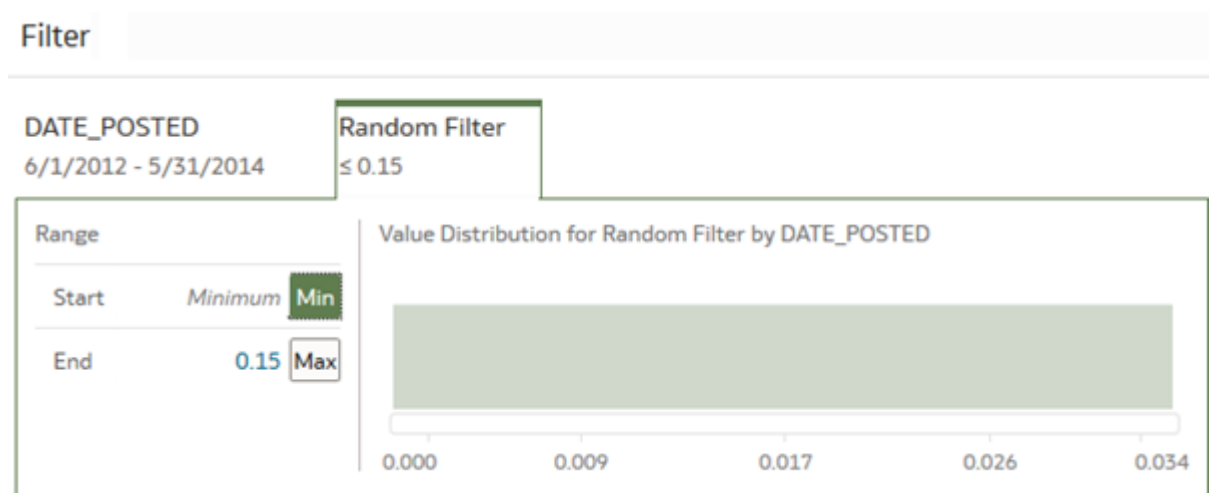
6. Click **Save**. In Save Workbook, enter Donations\_Workbook in **Name**, and then click **Save** .
7. Click **Go back** .


## Create a Random Dataset

1. On the Home page, click **Create**, and then click **Data Flow**.
2. In Add Data, select the donation dataset, and then click **Add**.
3. In Data Flow Steps, double-click **Filter**. In Filter, click **Add Filter** . From Available Data, select **DATE\_POSTED**. In Range values, enter 6/01/2012 in the first calendar text box. Enter 5/31/2014 in the second calendar text box. Click outside the dialog.



4. Click the **donation** node. Double-click **Add Columns**.
5. In Add Columns, enter Random Filter in **Name**, and then enter RAND() in the Expression field. Click **Validate**, and then click **Apply**.
6. Click the **Filter** node. From Available Data, click **Random Filter**. Click the **End** field, enter .15 to select a maximum of 15% of the sample data, and then click **End**.



7. Click **Add a step**  on the Filter node, and then click **Select Columns**. In Select Columns, select **Random Filter** from the Selected Columns list, and then click **Remove selected**.

The Random Filter column is no longer needed in the dataset.

donation → Add Columns → Filter → Select Columns

### Select Columns

Search  Add all Add selected Selected (26/27) Remove all Remove selected

Random Filter

- PROJECTID
- TEACHER\_ACCTID
- SCHOOL\_ID
- SCH\_LATITUDE
- SCH\_LONGITUDE
- SCH\_CITY
- SCH\_STATEZIP
- SCH\_METRO

PROJECTID	TEACHER_ACCTID	SCHOOL_ID	SCH_LATITUDE
238164	554050ab6859e8802e9b4cd0b0abae6c	6f1760717a868d936fe0acd06dfa8f13	40.746647
238188	4386a38a29a30ecfb5955d02374e6702	863d63cc4aa0b5542ea266c14e9fe6de	43.013488
238192	9a2a17fc6dc77f8b885e736c81e57287	eda3770dc6c610fedaec009b77ce9025	36.022288
238213	bacb8c2aaf1f569f27ed1ca56da1bd8f	4bbe5ba6764a80f088459229aaf3036a	39.790356
238227	a65f30ea26a6f9a2ba07f220c3eaf8ce	a58856103147b7c4db28fa2556add771	40.706825

- From Data Flow Steps, drag **Save Dataset** to the Select Columns node. In Save Dataset, enter sample\_donation\_data.
- Under Columns, in the PROJECTID row, select **Count** from Default Aggregation list.
- In the SCH\_LATITUDE row, click **Measure** and then select **Attribute**. In the SCH\_LONGITUDE row, click **Measure**, and then select **Attribute**.



donation
Add Columns
Filter
Select Columns
Save Data

### Save Dataset

Dataset

sample\_danation\_data

Dataset
Table

sample\_donation\_data

Description

Save data to


Dataset Storage

When Run



☐ Prompt to specify Dataset

Columns

Name	Treat As	Default Aggregation
PROJECTID	Measure	Count
TEACHER_ACCTID	Attribute	
SCHOOL_ID	Attribute	
SCH_LATITUDE	Measure	Sum
SCH_LONGITUDE	Measure	Sum
SCH_CITY	Attribute	
SCH_STATEZIP	Attribute	
SCH_METRO	Attribute	
SCH_COUNTY	Attribute	
SCH_CHARTER	Attribute	
TCHR_PREFIX	Attribute	
TCHR_TEACH_FOR_AMERICA	Attribute	

11. Click **Save**. In Save Data Flow As, enter sample\_donations\_data\_df, and then click **OK**.
12. Click **Run Data Flow**  to create the sample dataset.

## Examine the Sample Donations Dataset

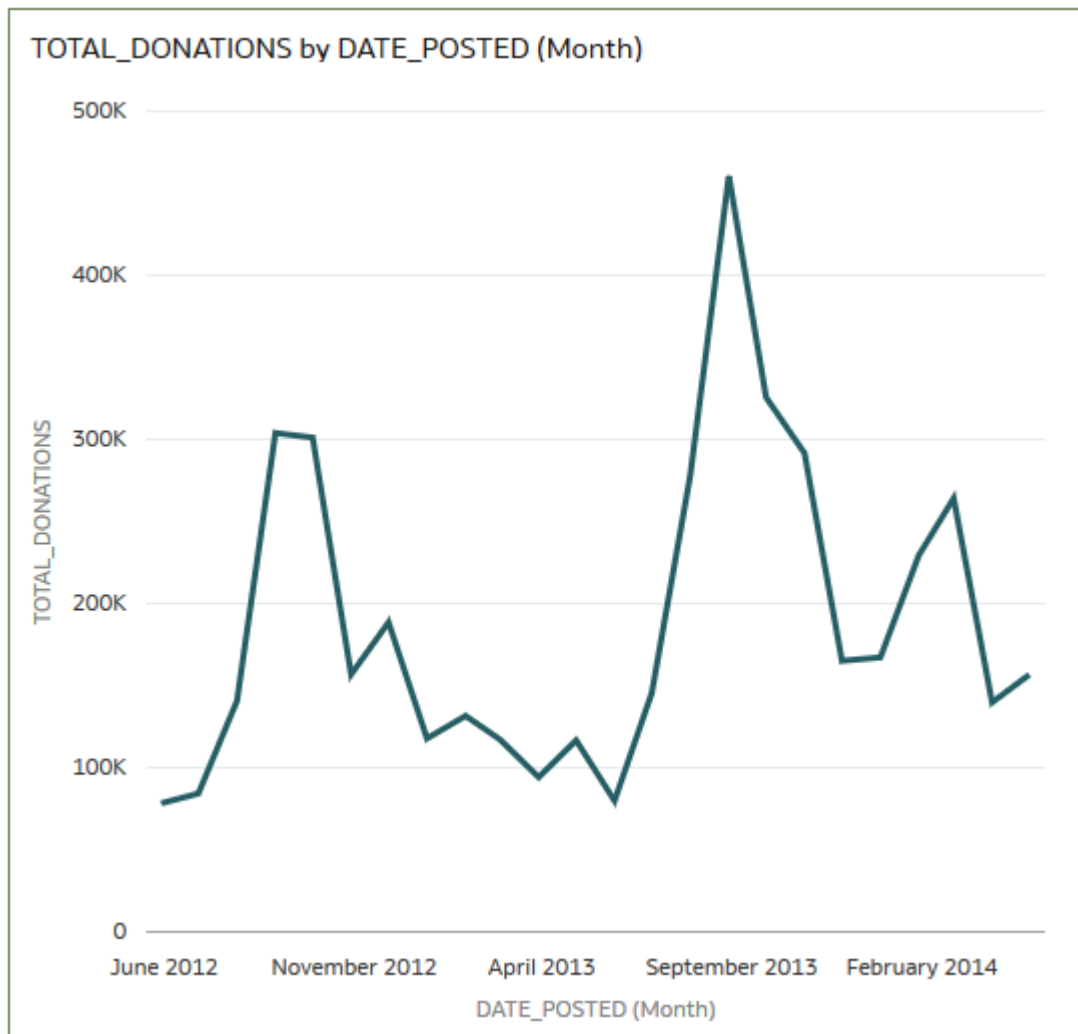
1. Click **Go back** . On the Home page, select the sample\_donation\_data dataset, click the **Actions menu** , and then select **Create Workbook**.
2. Right-click **PROJECTID**, select **Pick Visualization**, and then click **Tile**.


Because the sample data is a random selection of records from the dataset, your PROJECTID visualization might not match the results in this visualization.

PROJECTID

10,732

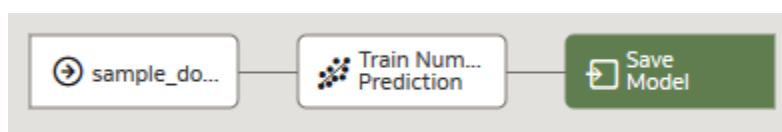
3. In the Data panel, hold down the **Ctrl** key, select **TOTAL\_DONATIONS** and **DATE\_POSTED**. Right-click and then select **Create the Best Visualization**.
4. In the Grammar panel, right-click **DATE\_POSTED**, select **Show By**, and then select **Month**.





- Click **Save**. In Save Workbook, enter donations\_random\_sample, and then click **Save**. Click **Go back**  to return to the Home page.

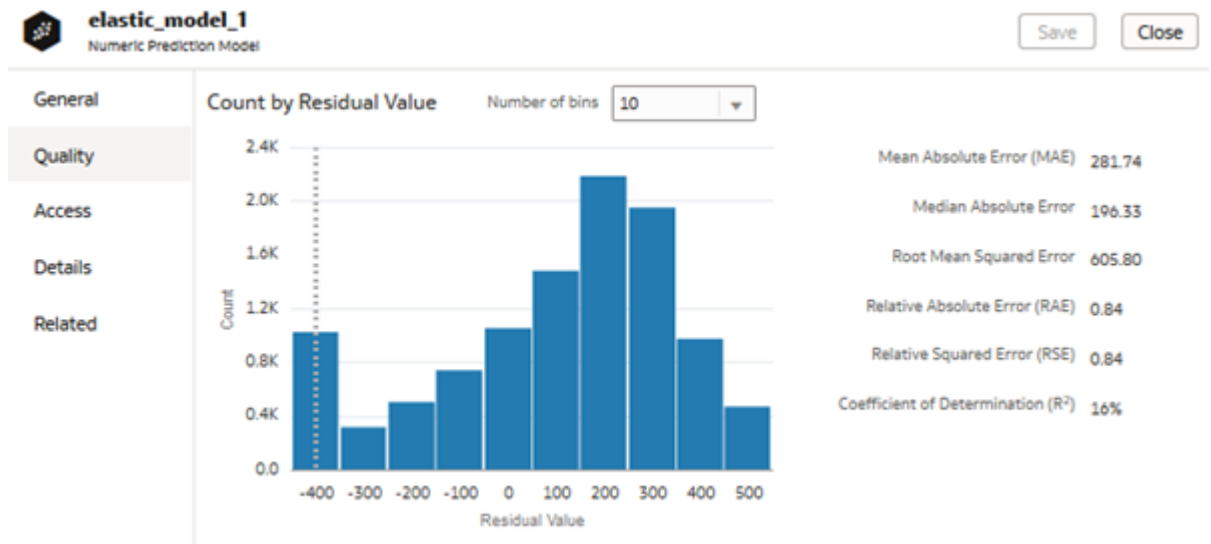
## Create a Training Model

- On the Home page, click **Create**, and then click **Data Flow**.
- In Add Data, select the sample\_donation\_data dataset, and then click **Add**.
- From Data Flow Steps, double-click **Train Numeric Prediction**.
- In Select Train Numeric Prediction Model Script, select **Elastic Net Linear Regression for model training**, and then click **OK**.
- In Train Numeric Prediction, click **Select a column**. From Available data, select **TOTAL\_DONATIONS** as the Target.
- Click the **Save Model** node in the data flow. Enter elastic\_model\_1 in **Model name**.






- Click **Save**. In Save Data Flow As, enter elastic\_train\_df in **Name**, and then click **OK**.
- Click **Run Data Flow**.

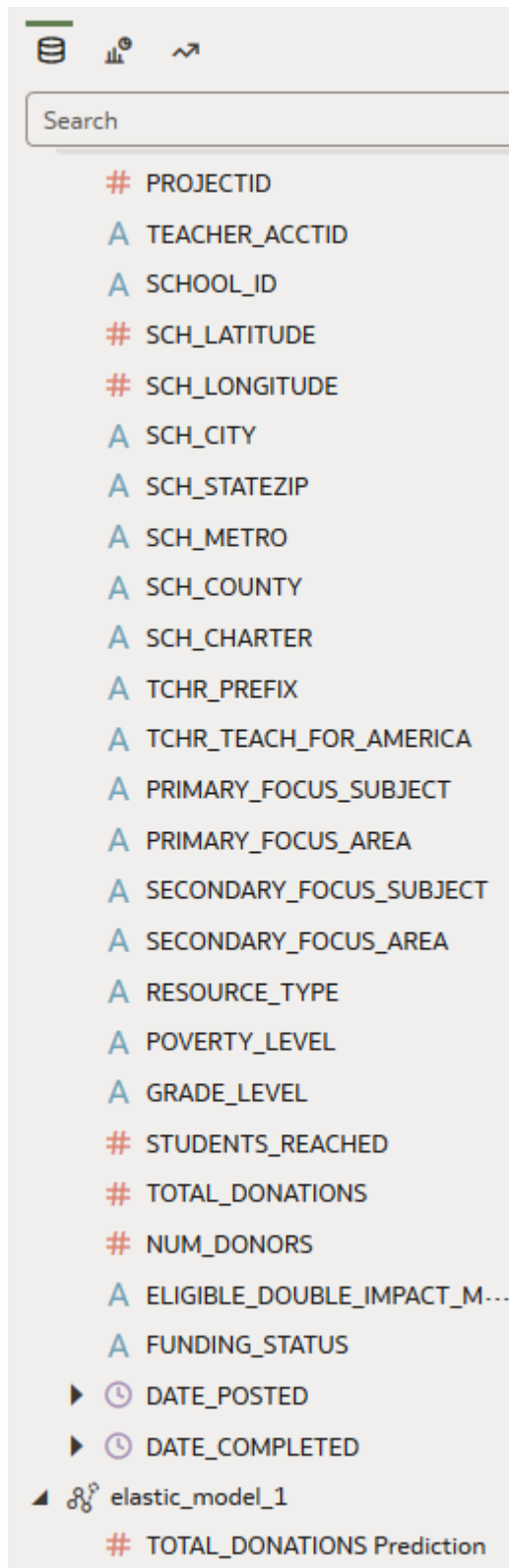
9. In the message "Data Flow elastic\_train\_df complete", click **Go back**  to return to the Home page.
10. On the Home page, click **Machine Learning** to view the elastic\_model\_1 output. Click the **Actions menu** , and then select **Inspect**.



## Apply the Train Model to a Workbook

In this section, you add the predicted value for total donations to the Total Donations by Date Posted (Month) visualization to view the results of using the elastic model.

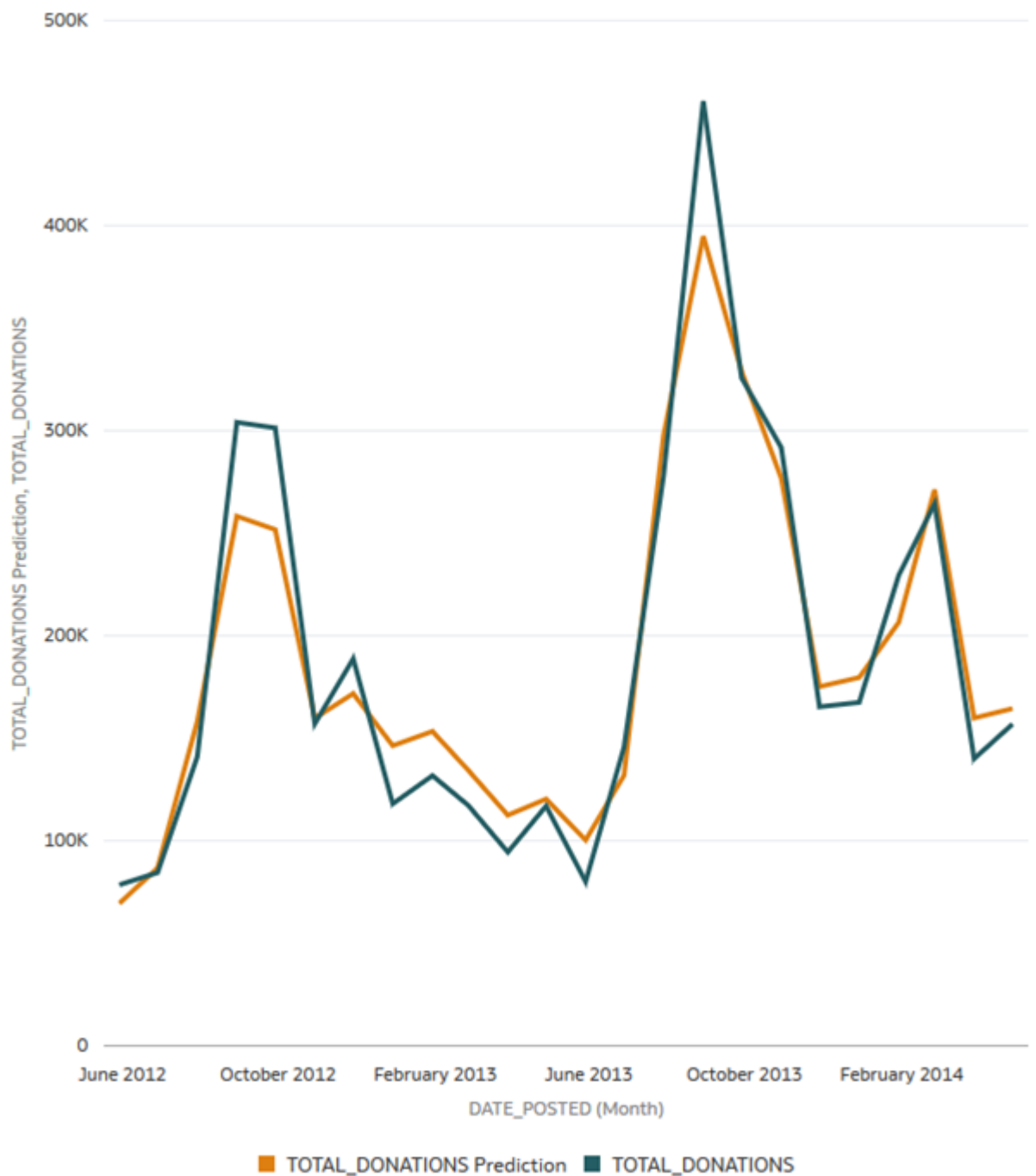
1. Click **Workbooks and Reports**.
2. On the Home page, search for your donations\_random\_sample workbook.
3. In the donations\_random\_sample workbook, click the **Actions Menu** , and then select **Open**.
4. In the PROJECTID visualization, click the **Menu** , and then select **Delete Visualization**.
5. In the Data panel, click **Add** , and then click **Create Scenario**.
6. In Create Scenario - Select Model, select **elastic\_model\_1**, and then click **OK**.



7. In the Data panel, expand **elastics\_model\_1**, select **TOTAL\_DONATIONS Prediction**, and then drag it to **Values (Y-Axis)** in the Grammar panel.

The green line represents the actual donations data by date posted. The orange line represents the predicted donations.

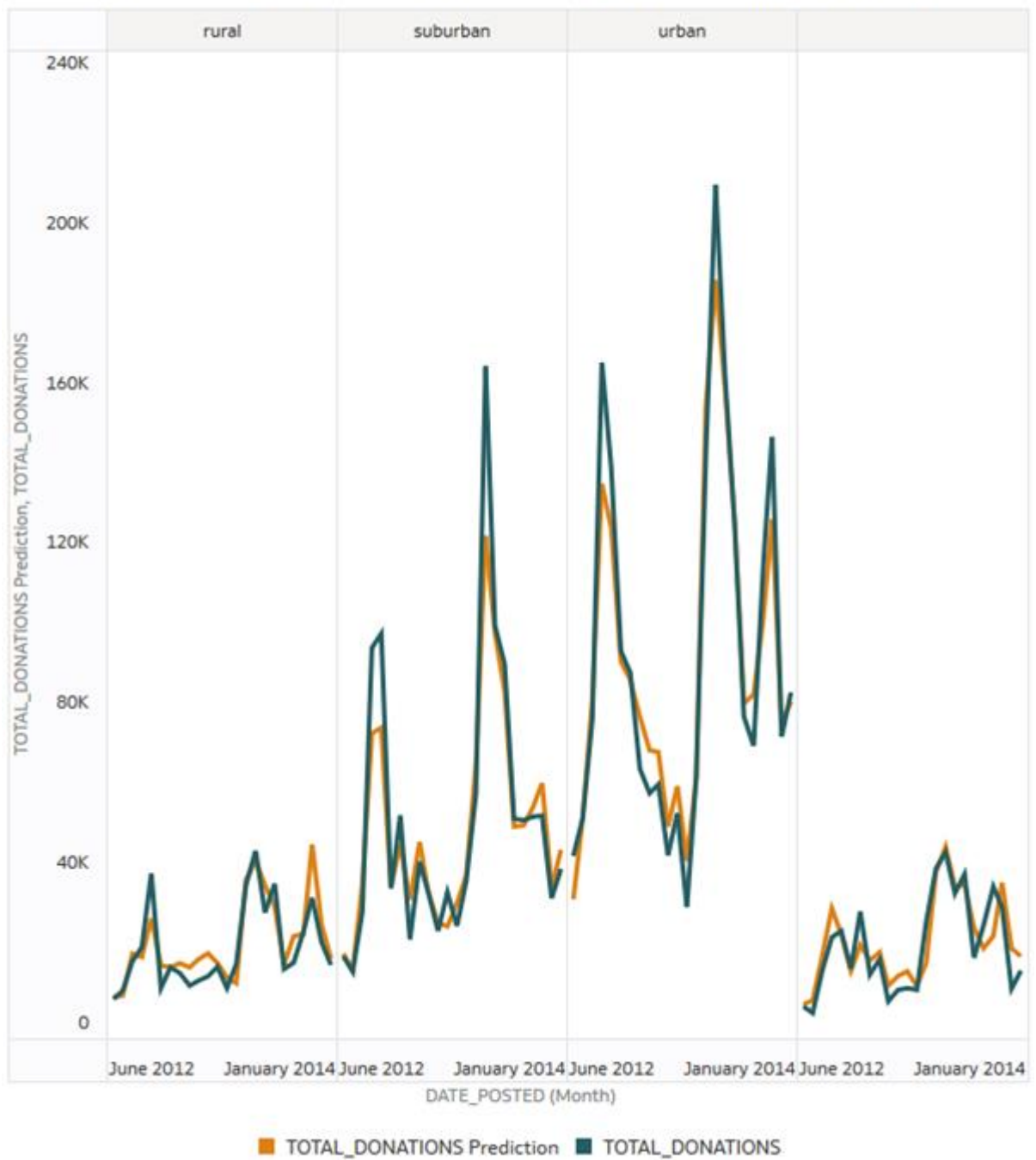
TOTAL\_DONATIONS Prediction, TOTAL\_DONATIONS by DATE\_POSTED (Month)



8. In the Data panel, select **SCH\_METRO**, drag it to **Trellis Columns** in the Grammar panel.

The visualization shows the donations data divided into school metro groups: rural, suburban, and urban.

TOTAL\_DONATIONS Prediction, TOTAL\_DONATIONS by DATE\_POSTED (Month), SCH\_METRO



- In the Grammar panel, click the **X** in **SCH\_METRO** to remove it from the visualization. Click **Save**.