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

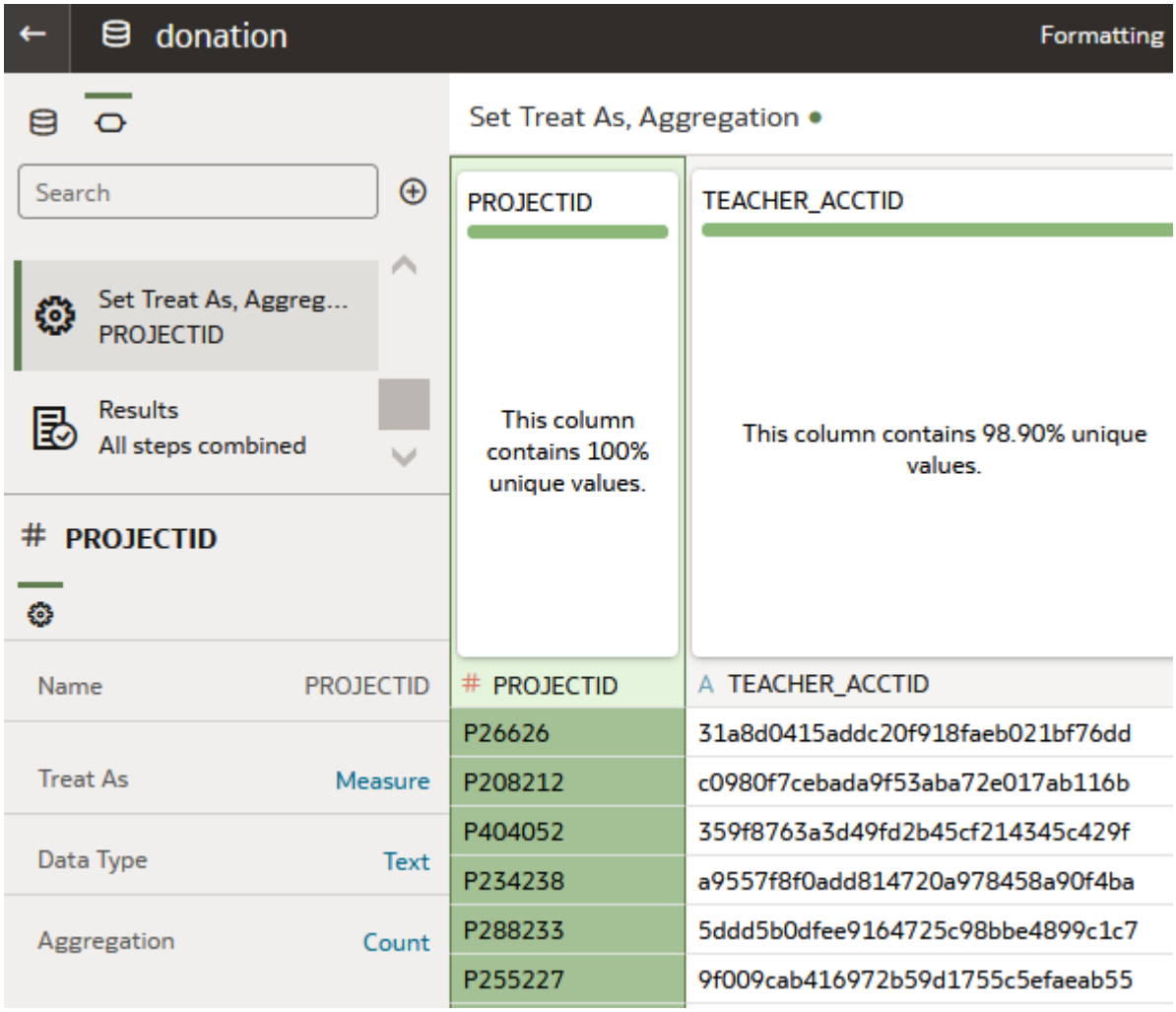
Create a Dataset

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**.

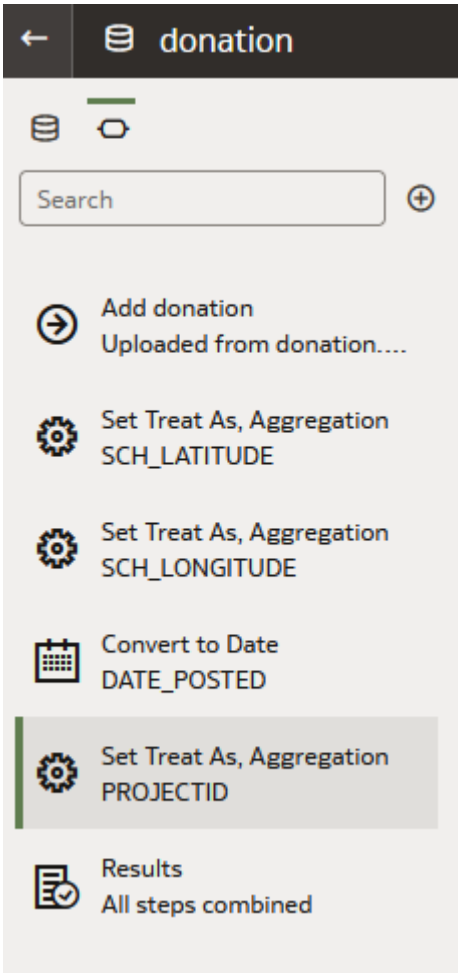
Prepare the Dataset

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



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.

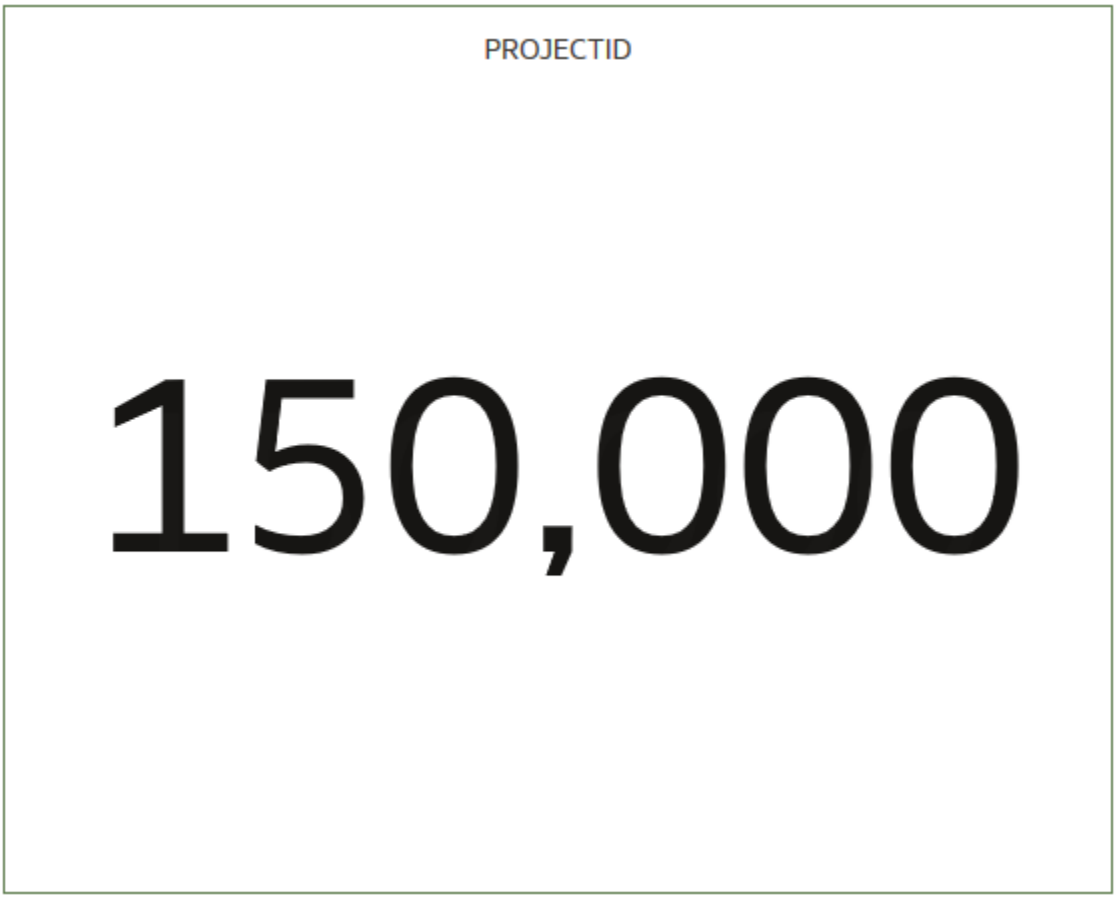


Visualize the Data

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.

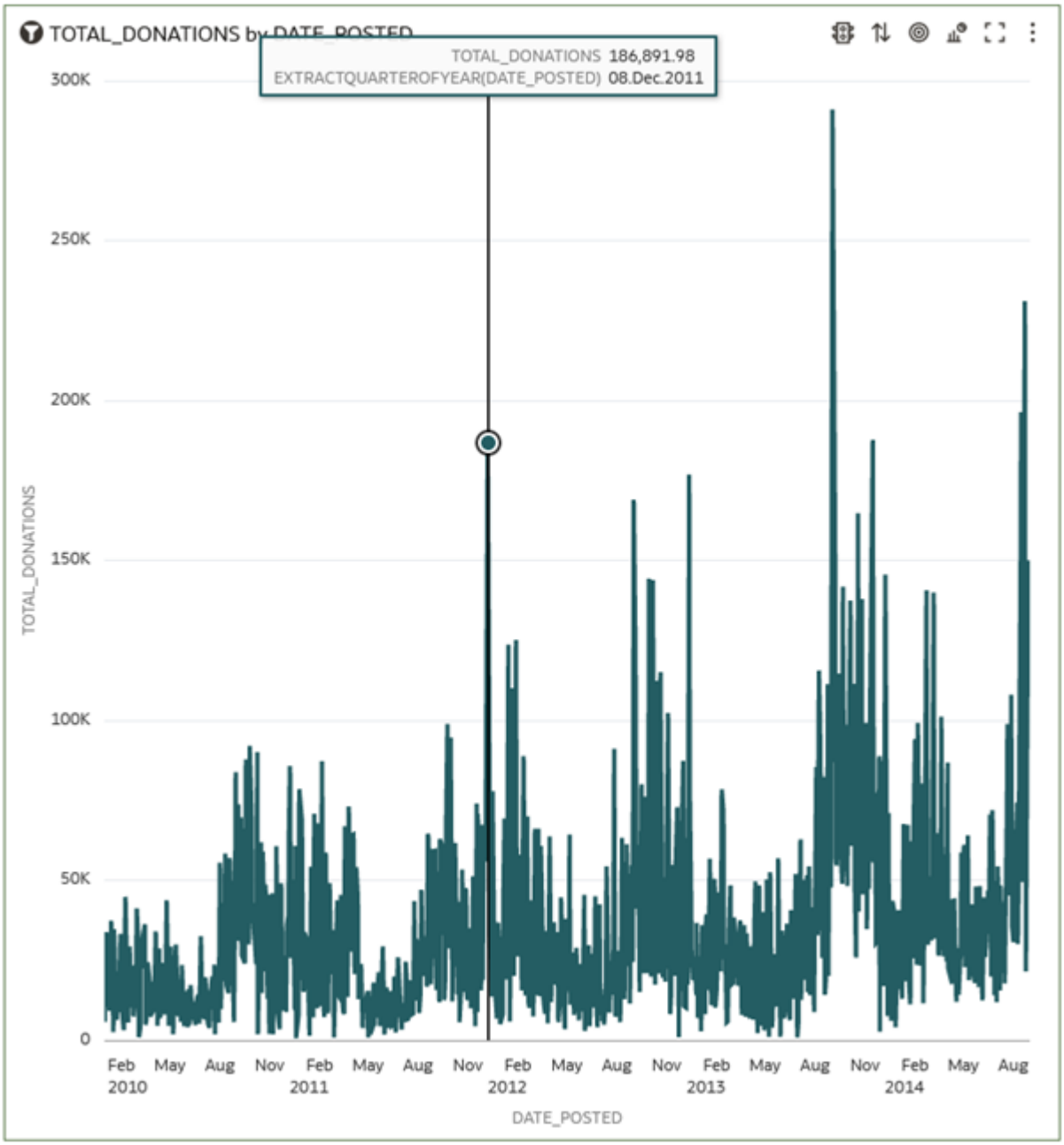
1. Click **Create Workbook**. Close the Auto Insights panel.
2. Drag **PROJECTID** from the Data panel, and then drop it on the visualization canvas.

   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**.

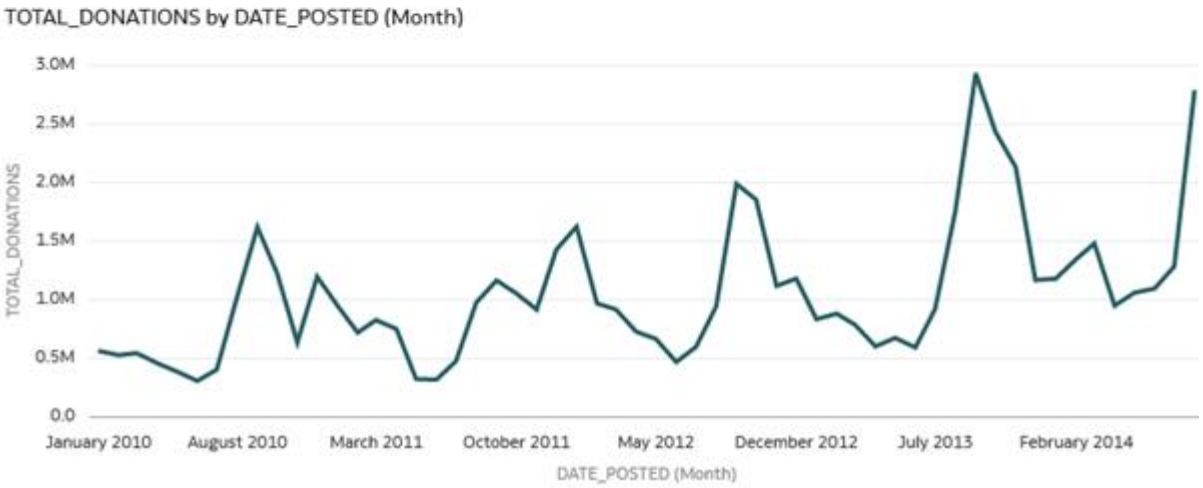
   In the line visualization, you can begin to see patterns in donations data.

TOTAL_DONATIONS

58,500,000.00

PROJECTID

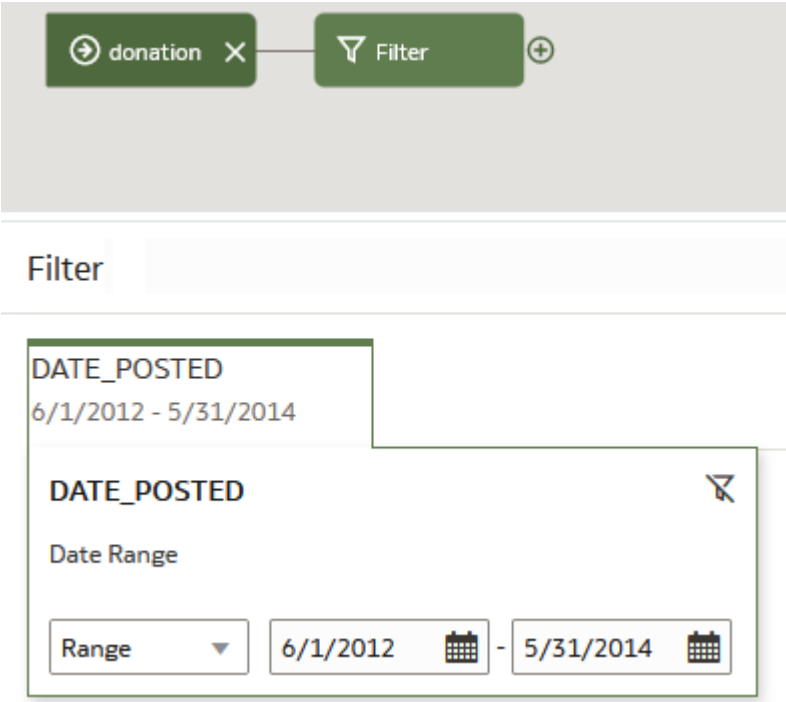
150,000



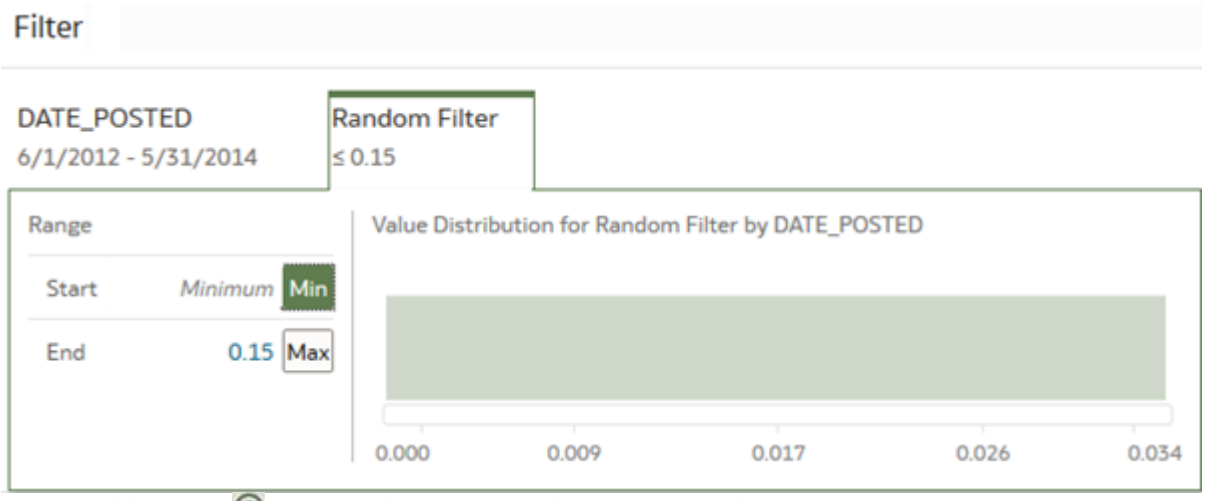
6. Click **Save**. In Save Workbook, enter Donations_Workbook in **Name**, and then click **Save**.
7. Click **Go back**.

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.

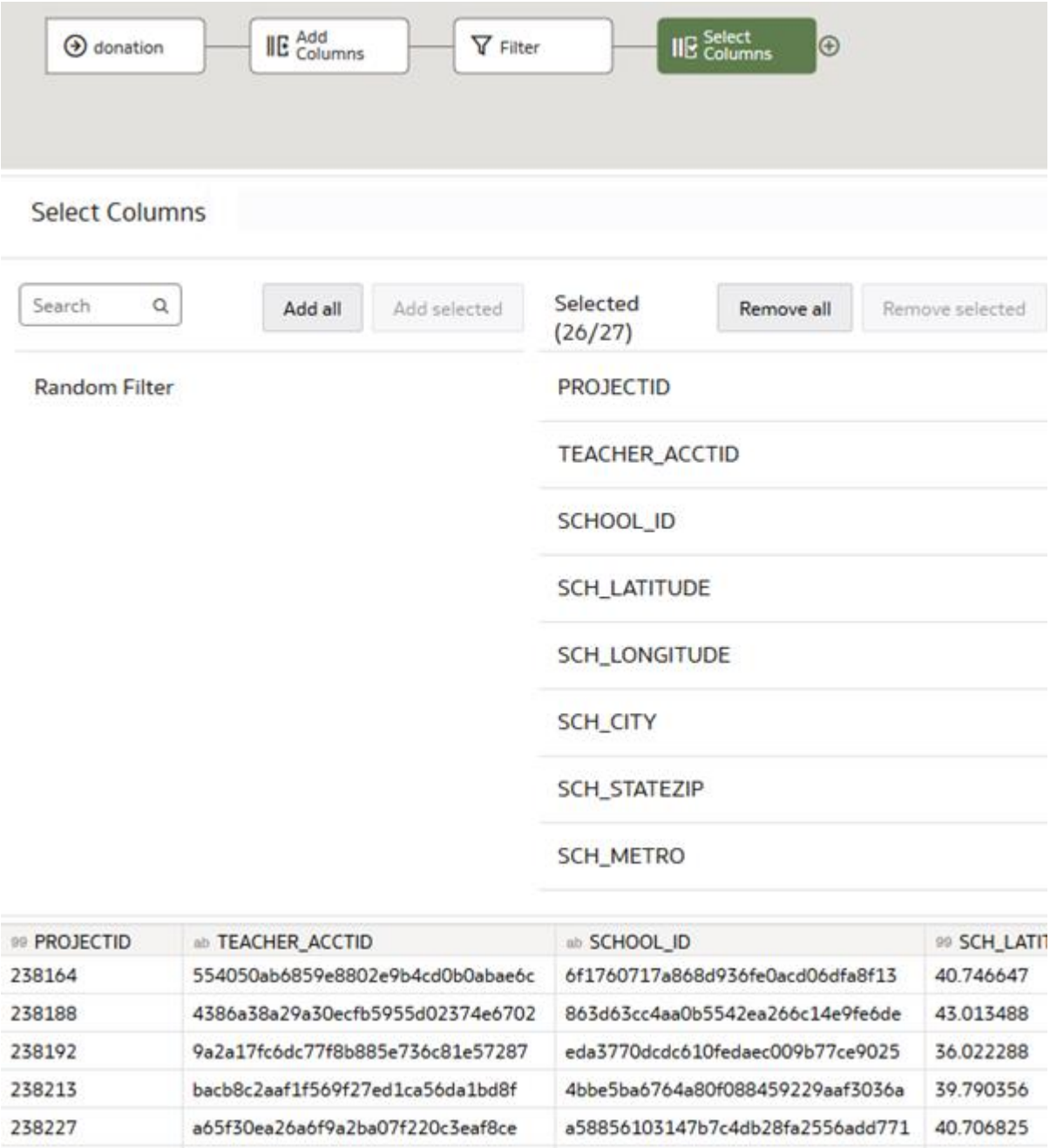


4. Click the **donation** node. Double-click **Add Columns**.
5. In Add Columns, enter Random Filter in **Name**, and then enter RAND() in the Expression field. Click **Validate**, and then click **Apply**.
6. Click the **Filter** node. From Available Data, click **Random Filter**. Click the **End**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

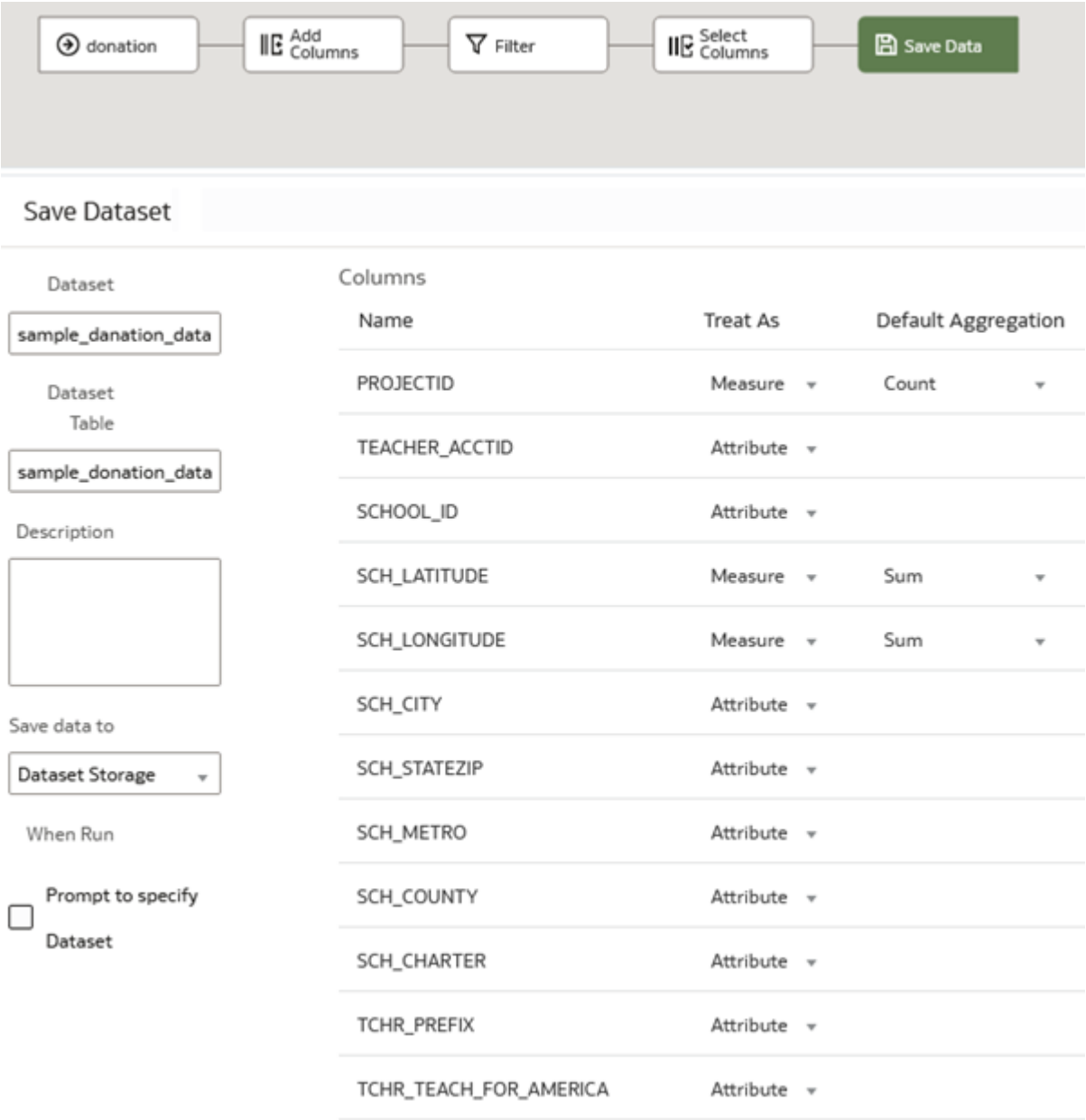


7. Click **Add a step** on the Filter node, and then click **Select Columns**. In Select Columns, select **Random Filter** from the Selected Columns list, and then click **Remove selected**.

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



8. From Data Flow Steps, drag **Save Dataset** to the Select Columns node. In Save Dataset, enter sample_donation_data.
9. Under Columns, in the PROJECTID row, select **Count** from Default Aggregation list.
10. In the SCH_LATITUDE row, click **Measure** and then select **Attribute**. In the SCH_LONGITUDE row, click **Measure**, and then select **Attribute**.

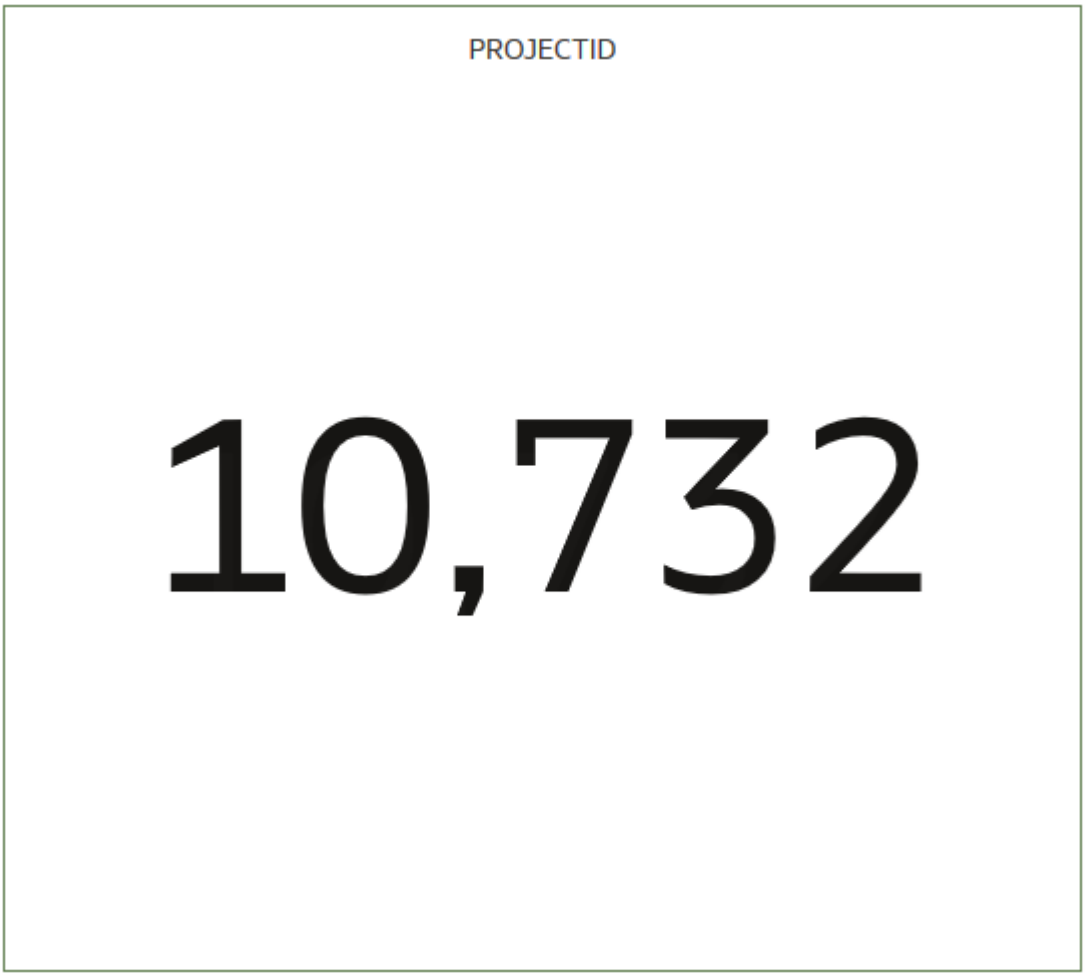


11. Click **Save**. In Save Data Flow As, enter sample_donations_data_df, and then click **OK**.
12. Click **Run Data Flow** to create the sample dataset.

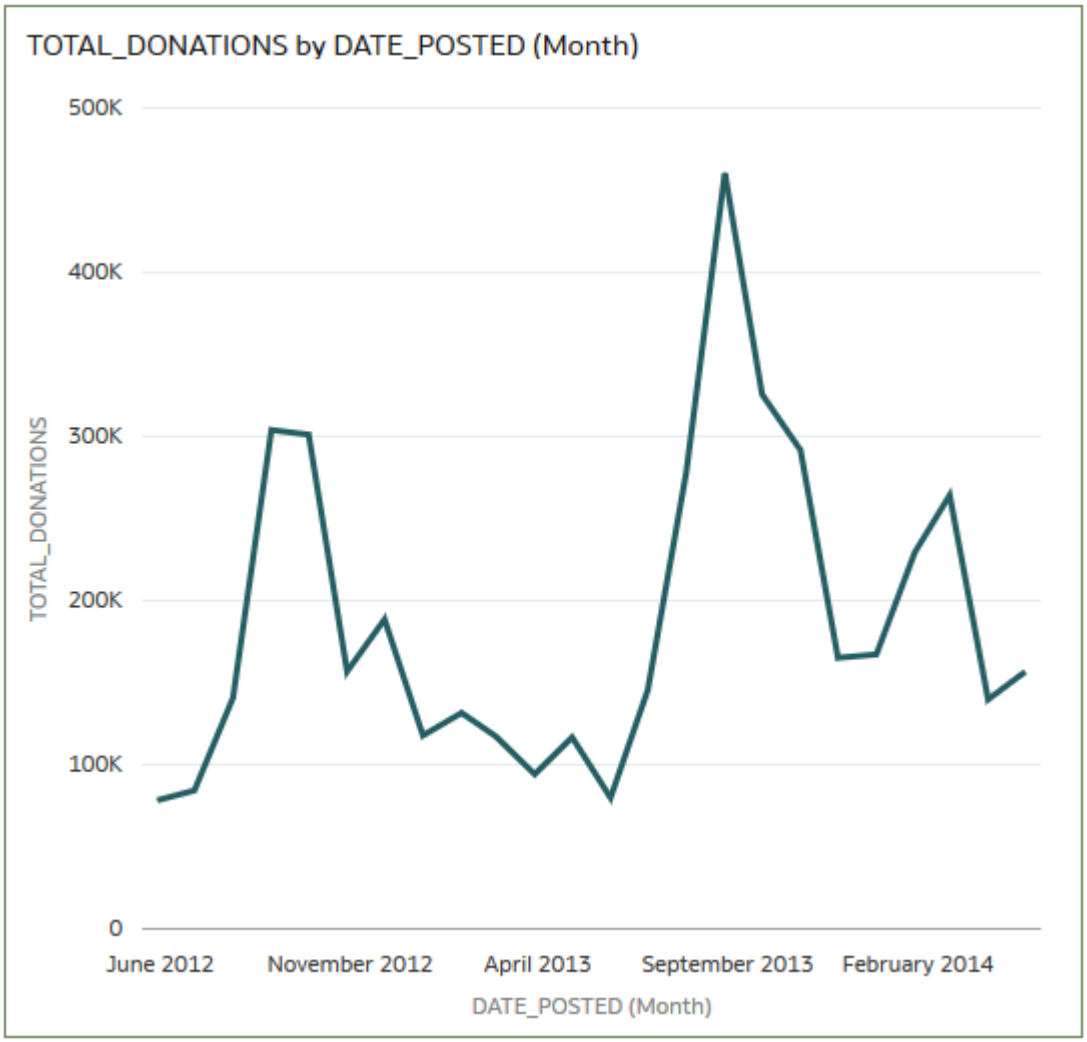
Examine the Sample Donations Dataset

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.



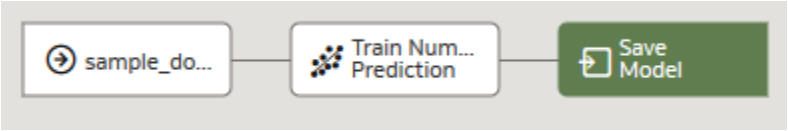
3. In the Data panel, hold down the **Ctrl** key, select **TOTAL_DONATIONS** and **DATE_POSTED**. Right-click and then select **Create the Best Visualization**.
4. In the Grammar panel, right-click **DATE_POSTED**, select **Show By**, and then select **Month**.



5. Click **Save**. In Save Workbook, enter donations_random_sample, and then click **Save**. Click **Go back** to return to the Home page.

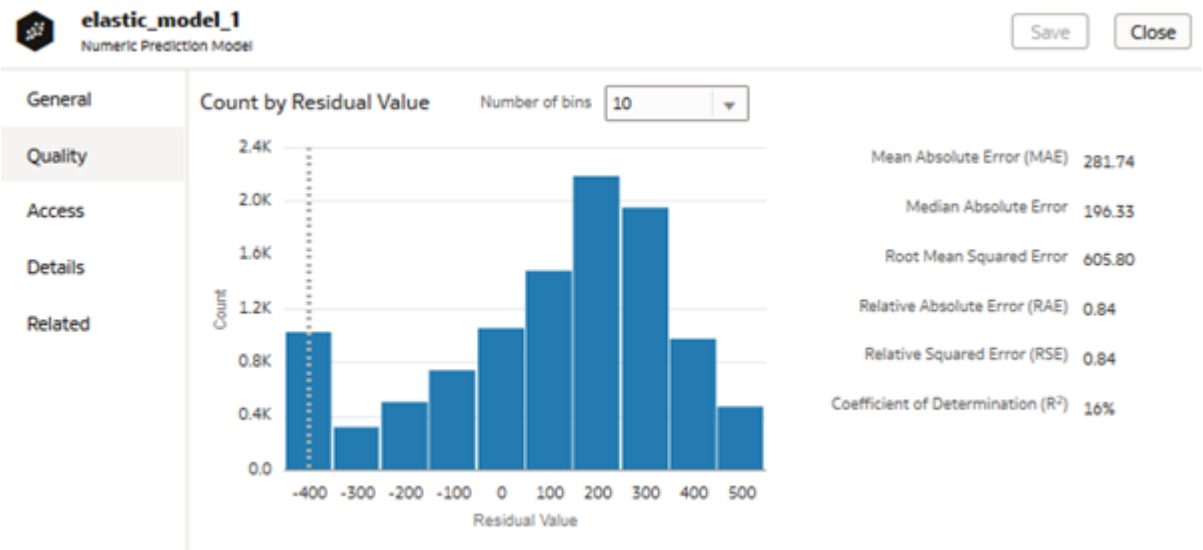
Create a Training Model

1. On the Home page, click **Create**, and then click **Data Flow**.
2. In Add Data, select the sample_donation_data dataset, and then click **Add**.
3. From Data Flow Steps, double-click **Train Numeric Prediction**.
4. In Select Train Numeric Prediction Model Script, select **Elastic Net Linear Regression for model training**, and then click **OK**.
5. In Train Numeric Prediction, click **Select a column**. From Available data, select **TOTAL_DONATIONS** as the Target.
6. Click the **Save Model** node in the data flow. Enter elastic_model_1 in **Model name.**



7. Click **Save**. In Save Data Flow As, enter elastic_train_df in **Name**, and then click **OK**.
8. Click **Run Data Flow**.

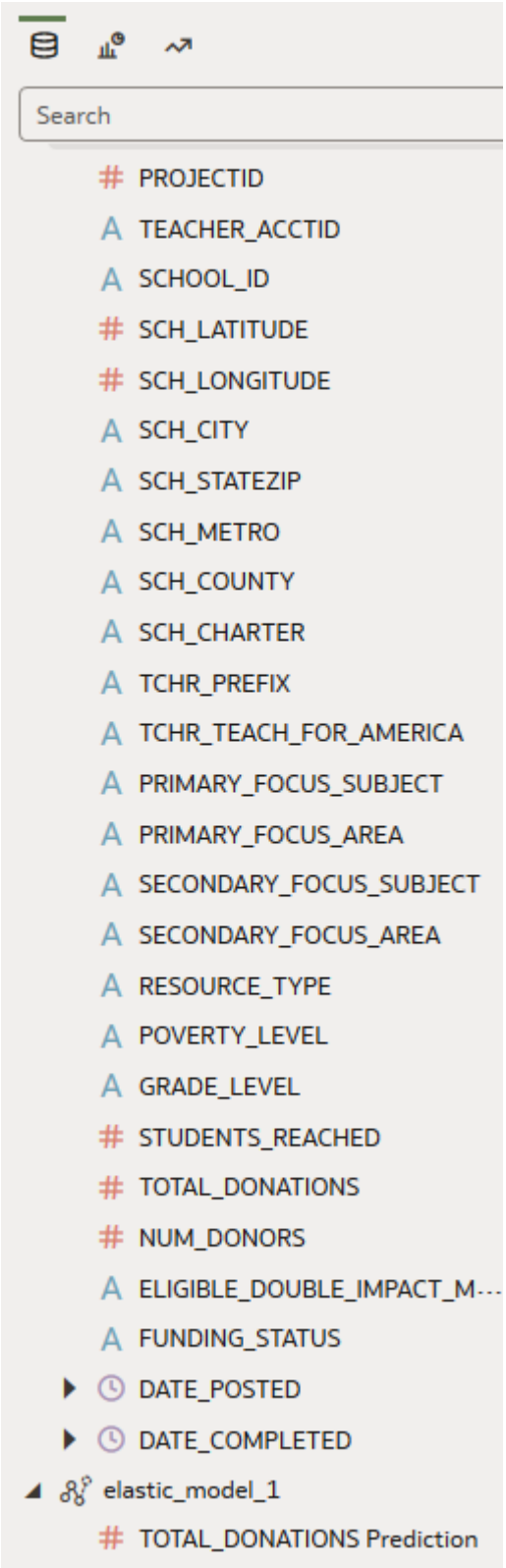
9. In the message "Data Flow elastic_train_df complete", click **Go back** to return to the Home page.
10. On the Home page, click **Machine Learning** to view the elastic_model_1 output. Click the **Actions menu**, and then select **Inspect**.



Apply the Train Model to a Workbook

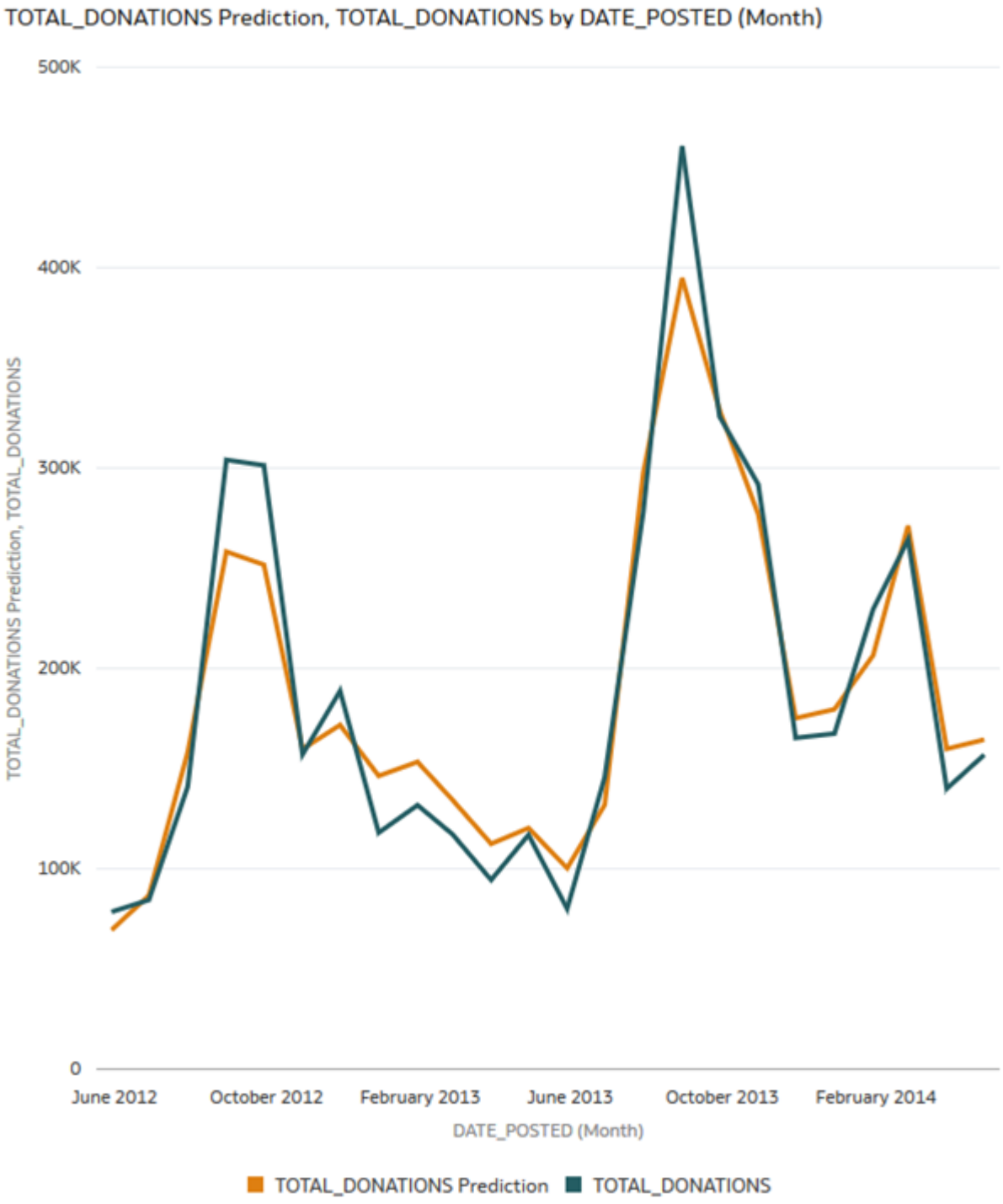
In this section, you add the predicted value for total donations to the Total Donations by Date Posted (Month) visualization to view the results of using the elastic model.

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**.



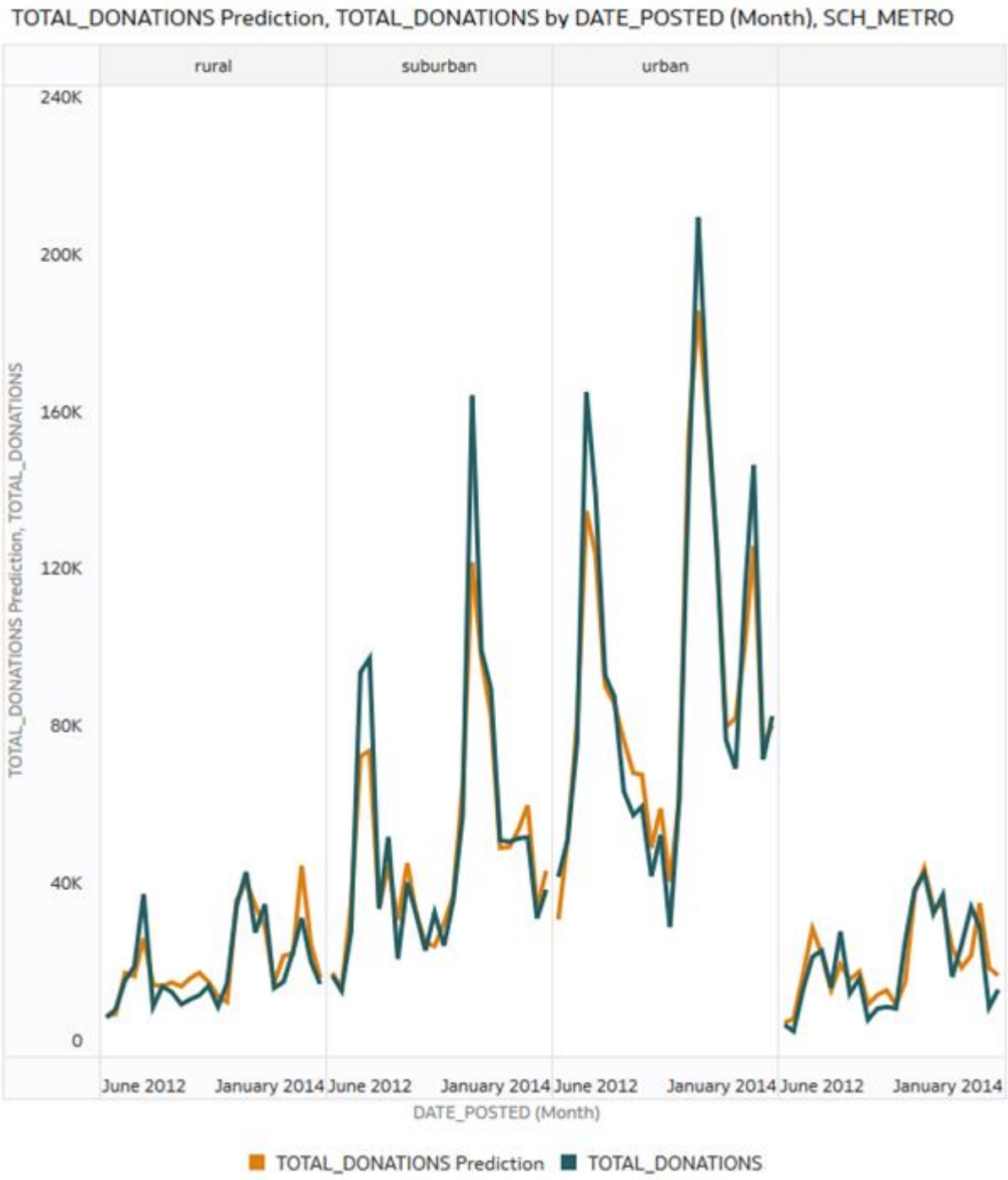
7. In the Data panel, expand **elastics_model_1**, select **TOTAL_DONATIONS Prediction**, and then drag it to **Values (Y-Axis)** in the Grammar panel.

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



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.



9. In the Grammar panel, click the **X** in **SCH_METRO** to remove it from the visualization. Click **Save**.