**Import, transform and export data**

Lab Overview

In this lab you learn how to import your own data in the designer to create custom solutions. There are two ways you can import data into the designer in Azure Machine Learning Studio:

* Azure Machine Learning datasets

Register datasets in Azure Machine Learning to enable advanced features that help you manage your data.

* Import Data module

Use the Import Data module to directly access data from online datasources.

The first approach will be covered later in the [next lab](https://udacity-manage.cloudlabs.ai/lab-03/README.md), which focuses on registering and versioning a dataset in Azure Machine Learning studio.

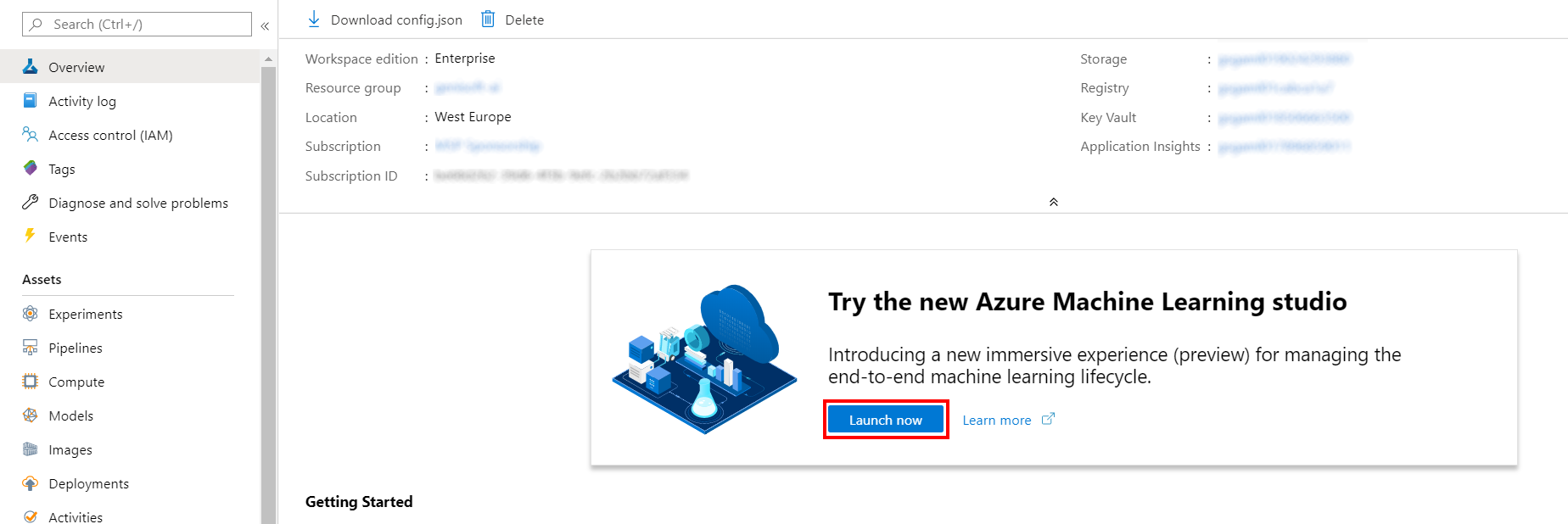
While the use of datasets is recommended to import data, you can also use the Import Data module from the designer. Data comes into the designer from either a Datastore or from Tabular Datasets. Datastores will be covered later in this course, but just for a quick definition, you can use Datastores to access your storage without having to hard code connection information in your scripts. As for the second option, the Tabular datasets, the following datasources are supported in the designer: Delimited files, JSON files, Parquet files or SQL queries.

The following exercise focuses on the Import Data module to load data into a machine learning pipeline from several datasets that will be merged and restructured. We will be using some sample data from the UCI dataset repository to demonstrate how you can perform basic data import transformation steps with the modules available in Azure Machine Learning designer.

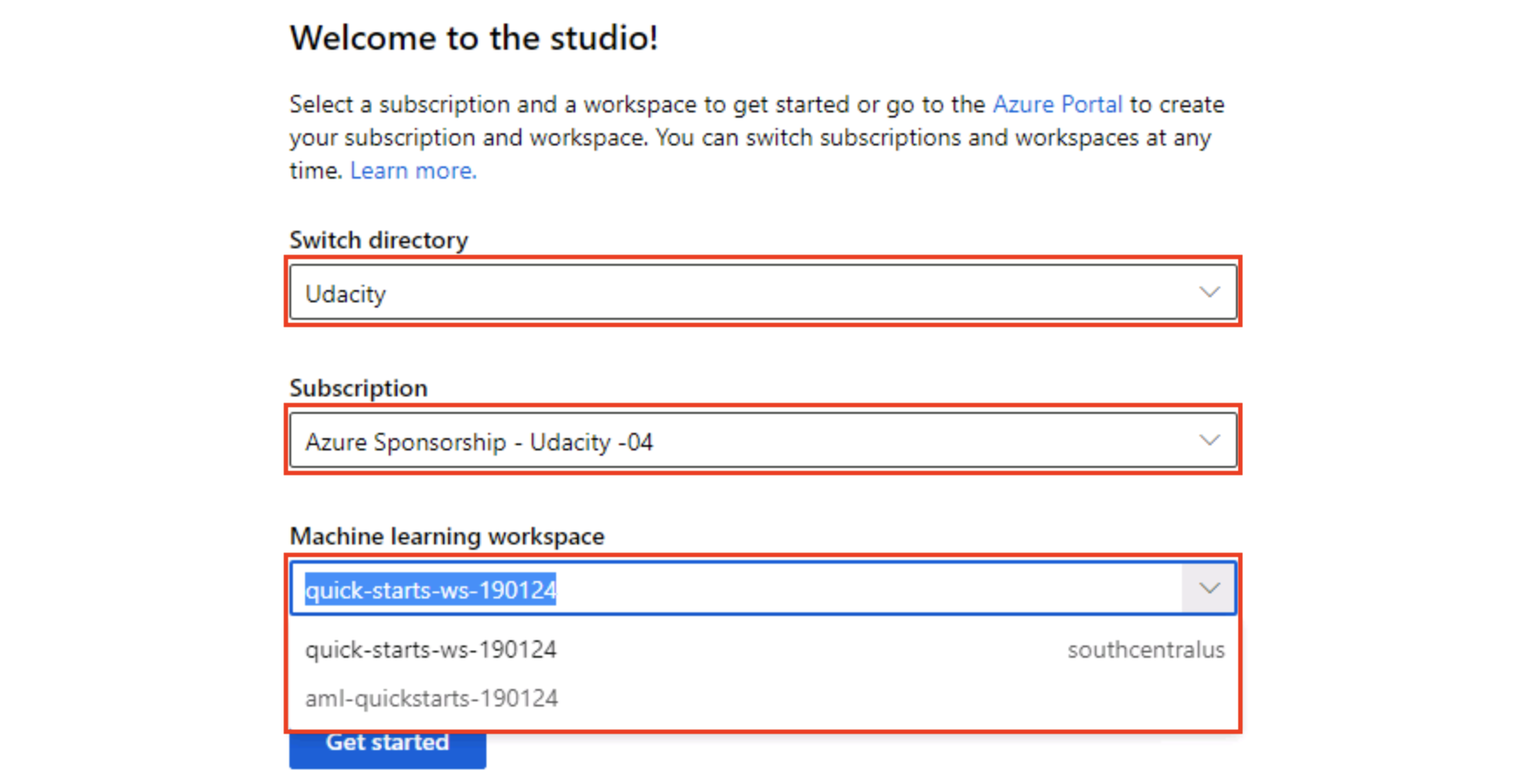
Exercise 1: Import, transform and export data using the Visual Pipeline Authoring Editor

Task 1: Open Pipeline Authoring Editor

1. In [Azure portal](https://portal.azure.com/), open the available machine learning workspace.
2. Select **Launch now** under the **Try the new Azure Machine Learning studio** message.

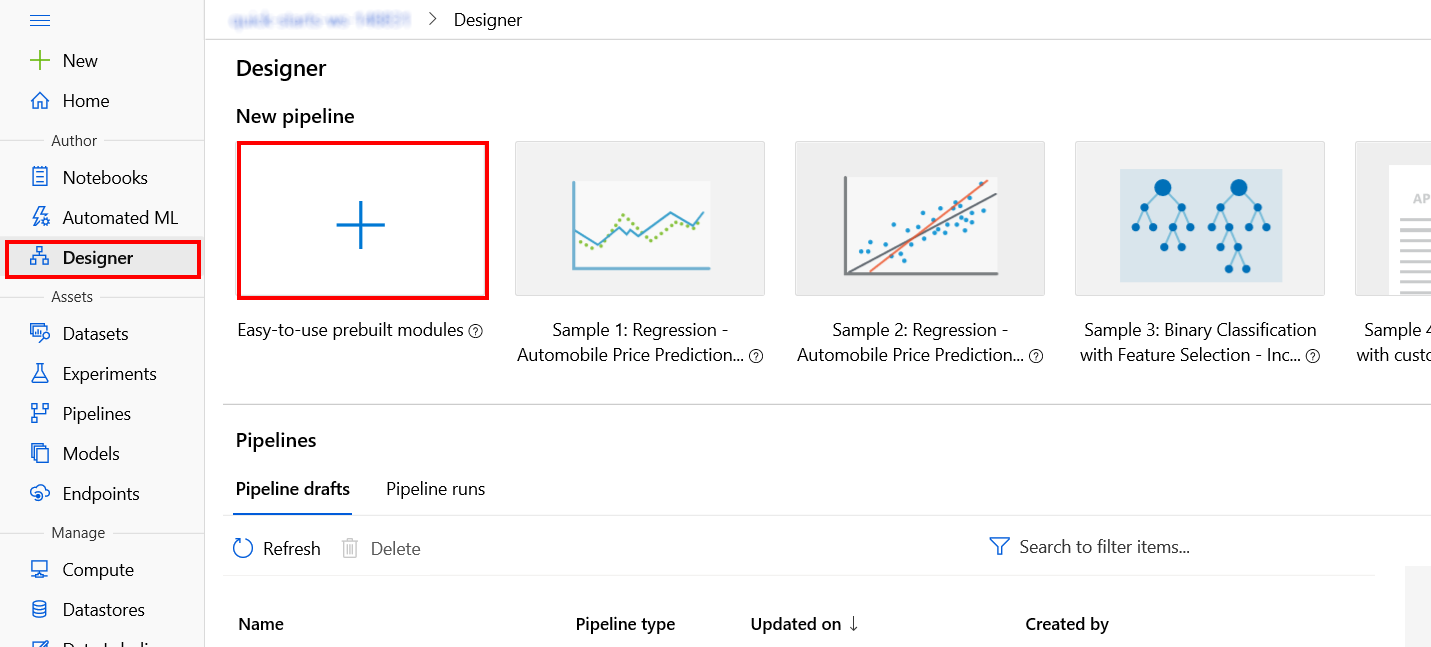


1. When you first launch the studio, you may need to set the directory and subscription. If so, you will see this screen:



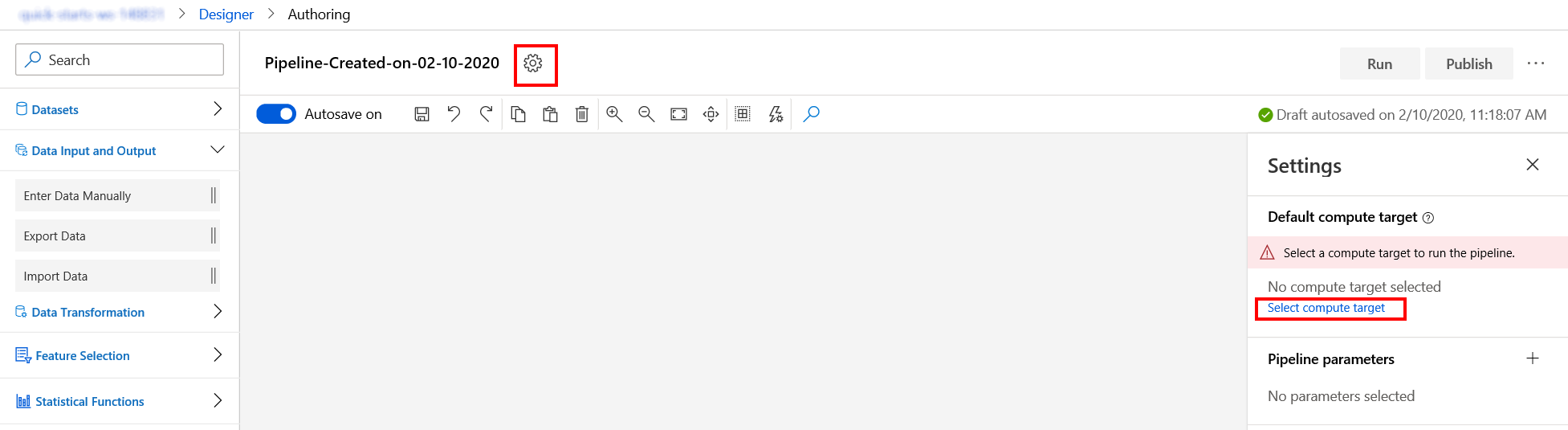
For the directory, select **Udacity** and for the subscription, select **Azure Sponsorship**. For the machine learning workspace, you may see multiple options listed. **Select any of these** (it doesn’t matter which) and then click **Get started**.

1. From the studio, select **Designer, +**. This will open a visual pipeline authoring editor.



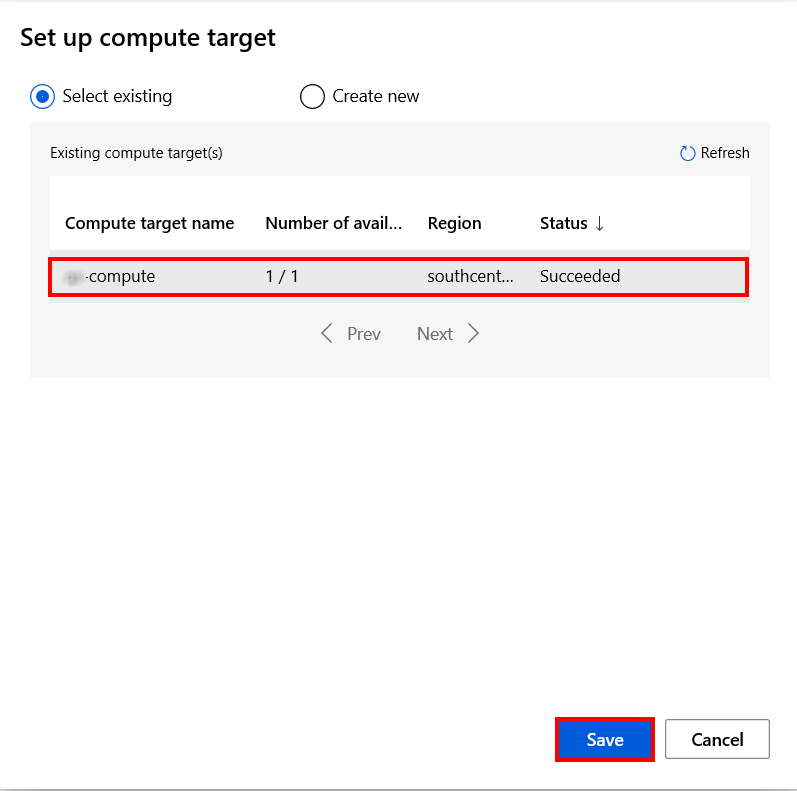
Task 2: Setup Compute Target

1. In the settings panel on the right, select **Select compute target**.



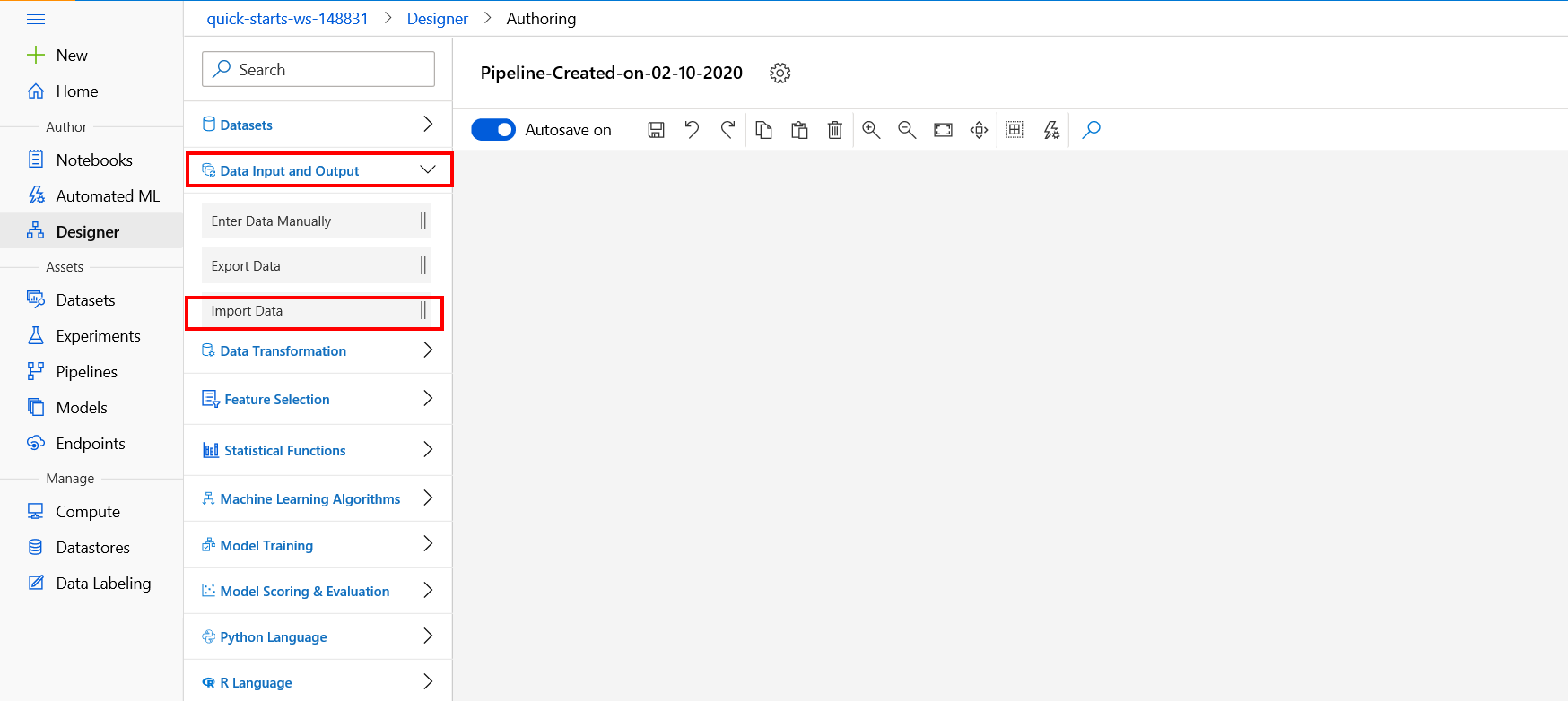
1. In the Set up compute target editor, select the existing compute target, and then select **Save**.

Note: If you are facing difficulties in accessing pop-up windows or buttons in the user interface, please refer to the Help section in the lab environment.

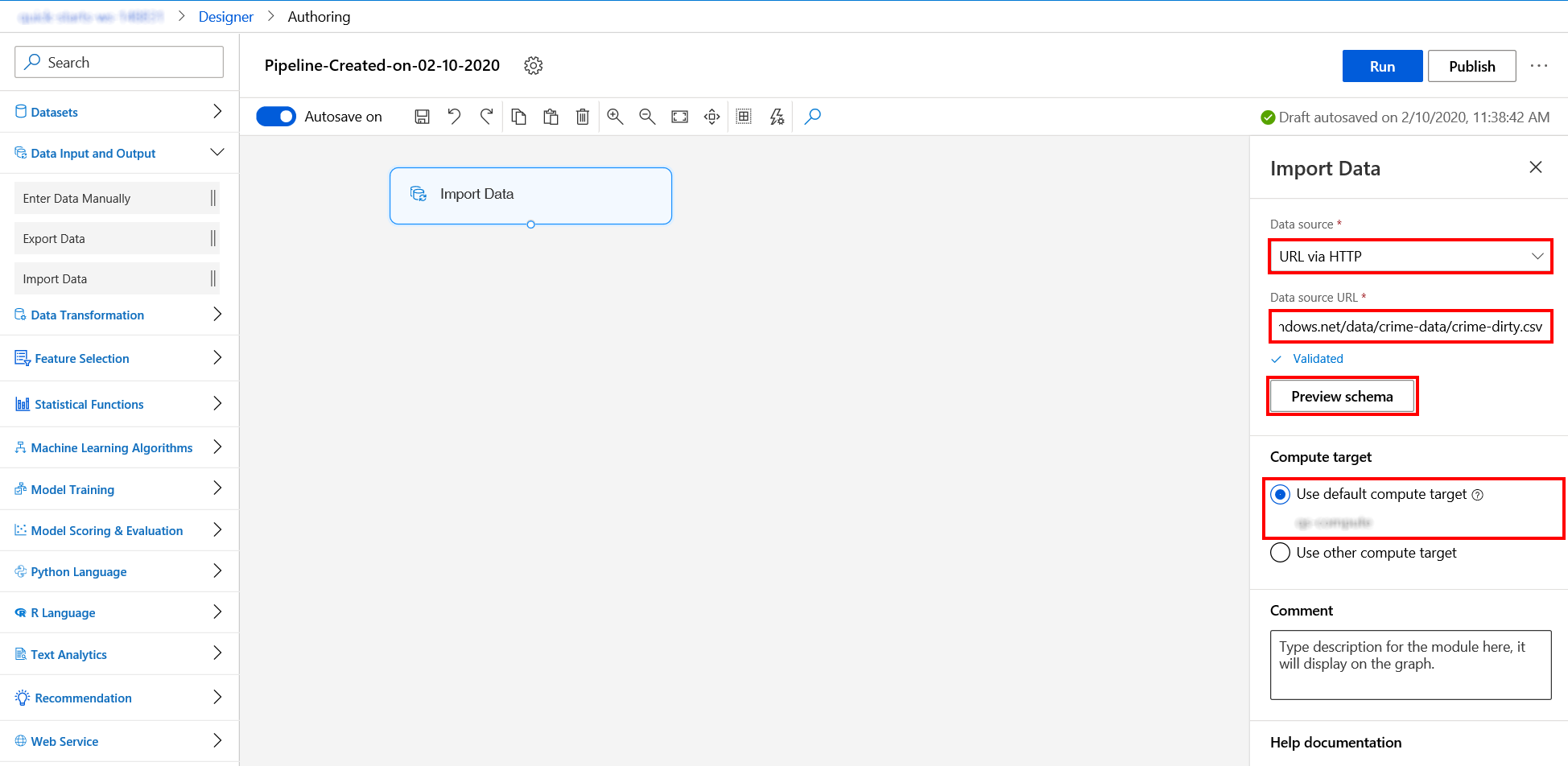


Task 3: Import data from Web URL

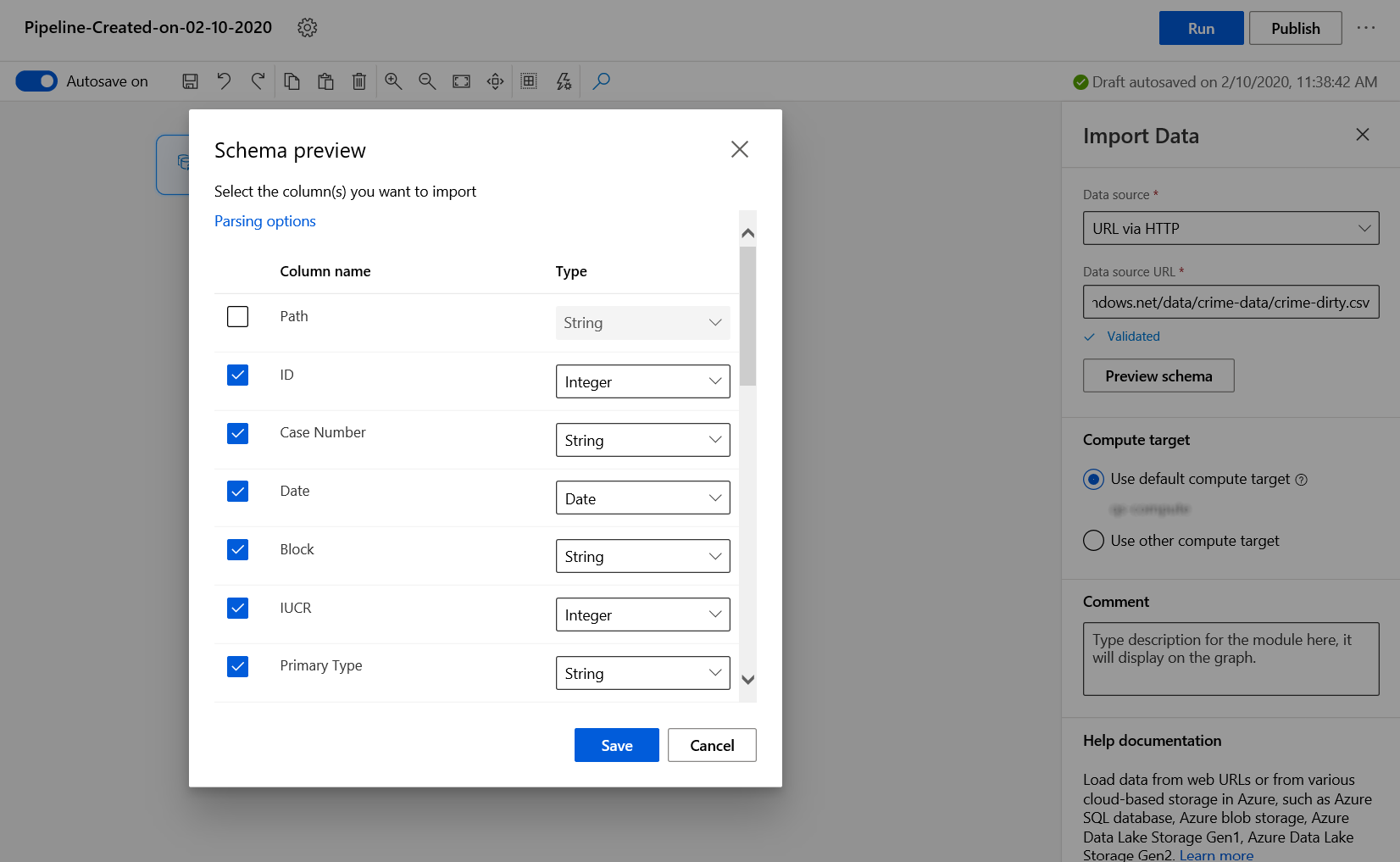
1. Select **Data Input and Output** section in the left navigation. Next, select **Import Data** and drag and drop the selected module on to the canvas.



1. In the Import data panel on the right, select the **URL via HTTP** option in the Data Source drop-down and provide the following Data source URL for the first CSV file you will import in your pipeline: https://introtomlsampledata.blob.core.windows.net/data/crime-data/crime-dirty.csv

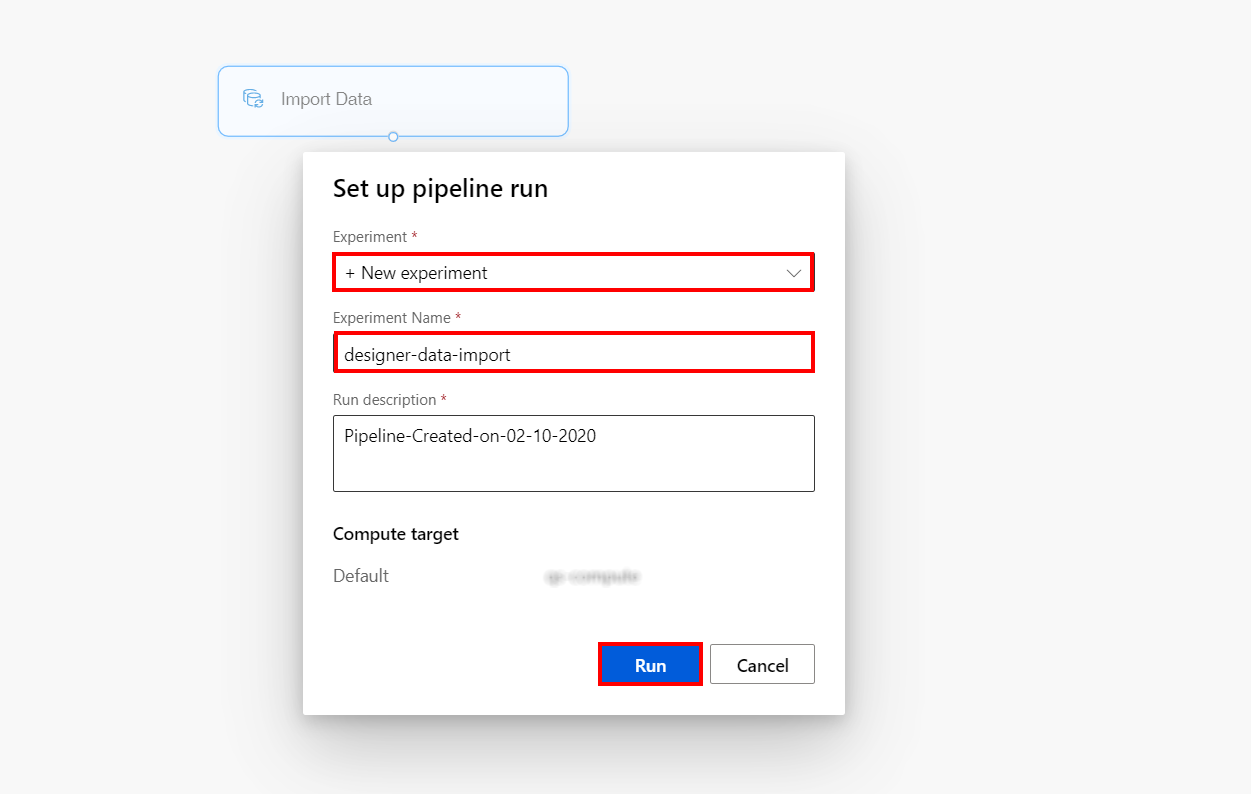


1. Select the **Preview schema** to filter the columns you want to include. You can also define advanced settings like Delimiter in Parsing options. Select **Save** to close the dialog.



Task 4: Create Experiment and Submit Pipeline

1. Back to the pipeline canvas, select **Submit** on the top right corner to open the Setup pipeline run editor.
2. In the Setup pipeline run editor, select **Experiment, Create new** and provide New experiment name: **designer-data-import**, and then select **Submit**.

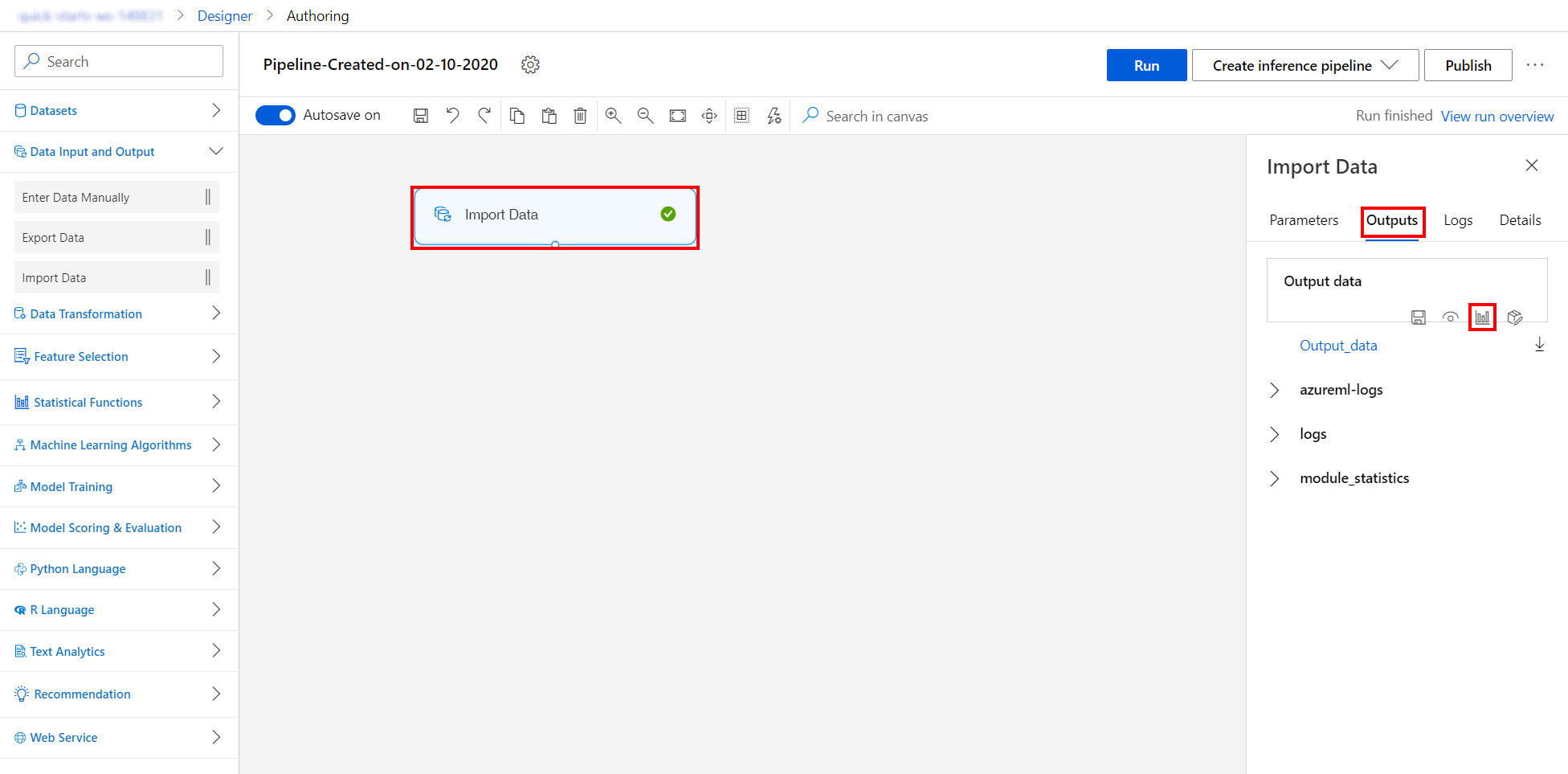


Please note that the button name in the UI is changed from **Run** to **Submit**.

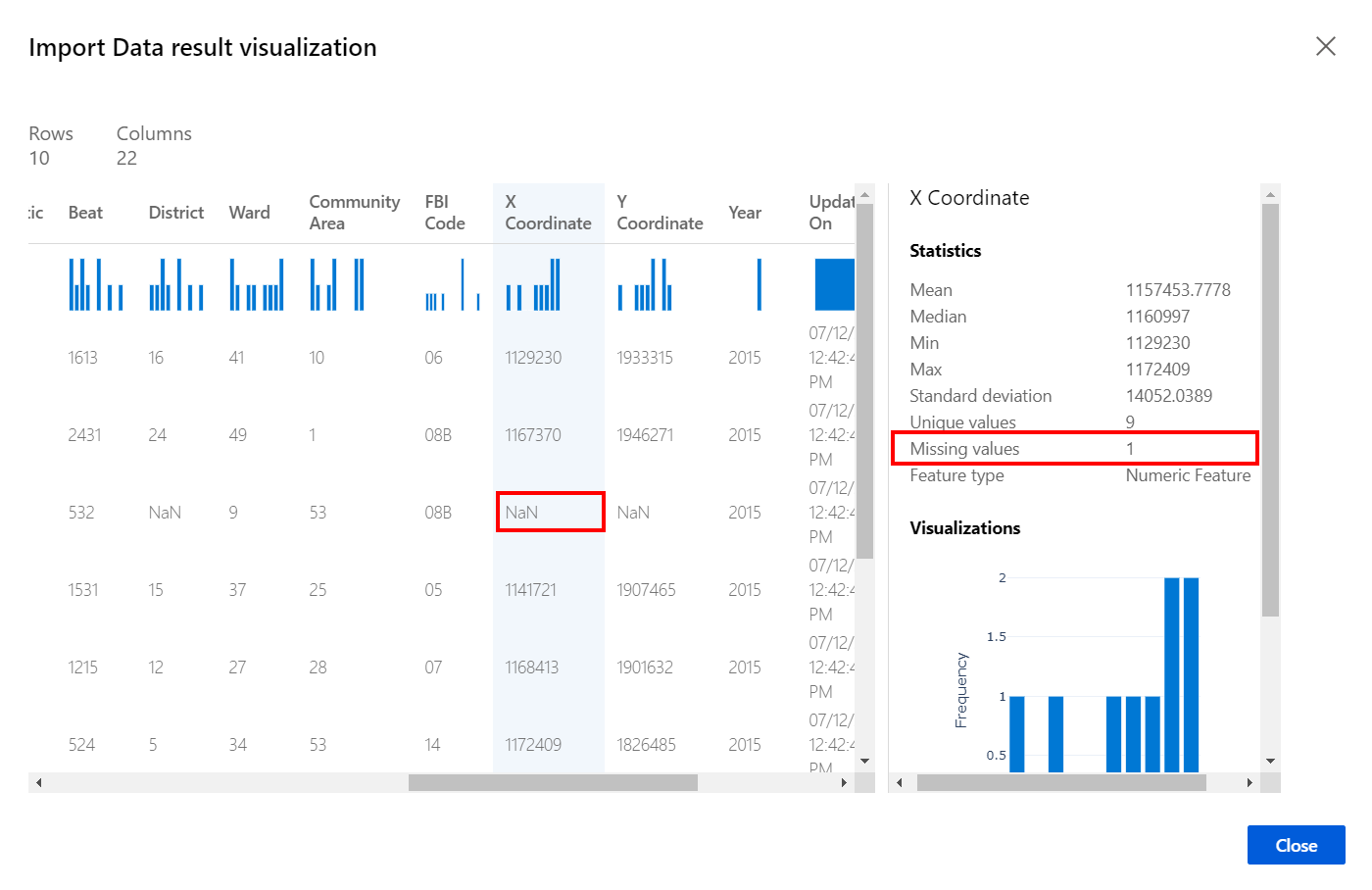
1. Wait for pipeline run to complete. It will take around **10 minutes** to complete the run.

Task 5: Visualize Import Data results

1. Select the Import Data module on the canvas and then select **Outputs** on the right pane. Click on the **Visualize** icon to open the Import Data result visualization dialog.



1. In the Import Data result visualization dialog take some moments to explore all the metadata that is now available to you, such as: number of rows, columns, preview of data and for each column you select you can observe: **Mean**, **Median**, **Min**, **Max** and also number of **Unique Values** and **Missing Values**. Data profiles help you glimpse into the column types and summary statistics of a dataset. Scroll right and select the **X Coordinate** column. Notice the Nan value on the third row in the preview table and check the Missing values number in the **Statistics** section.

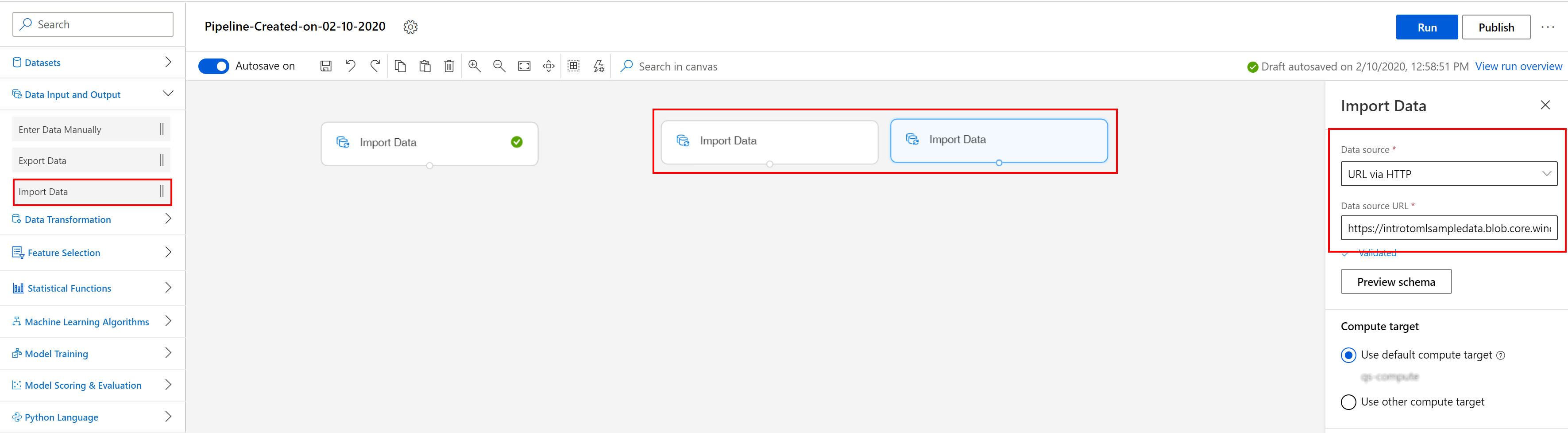


1. Select **Close** to return to the pipeline designer canvas where you can continue the data import phase.

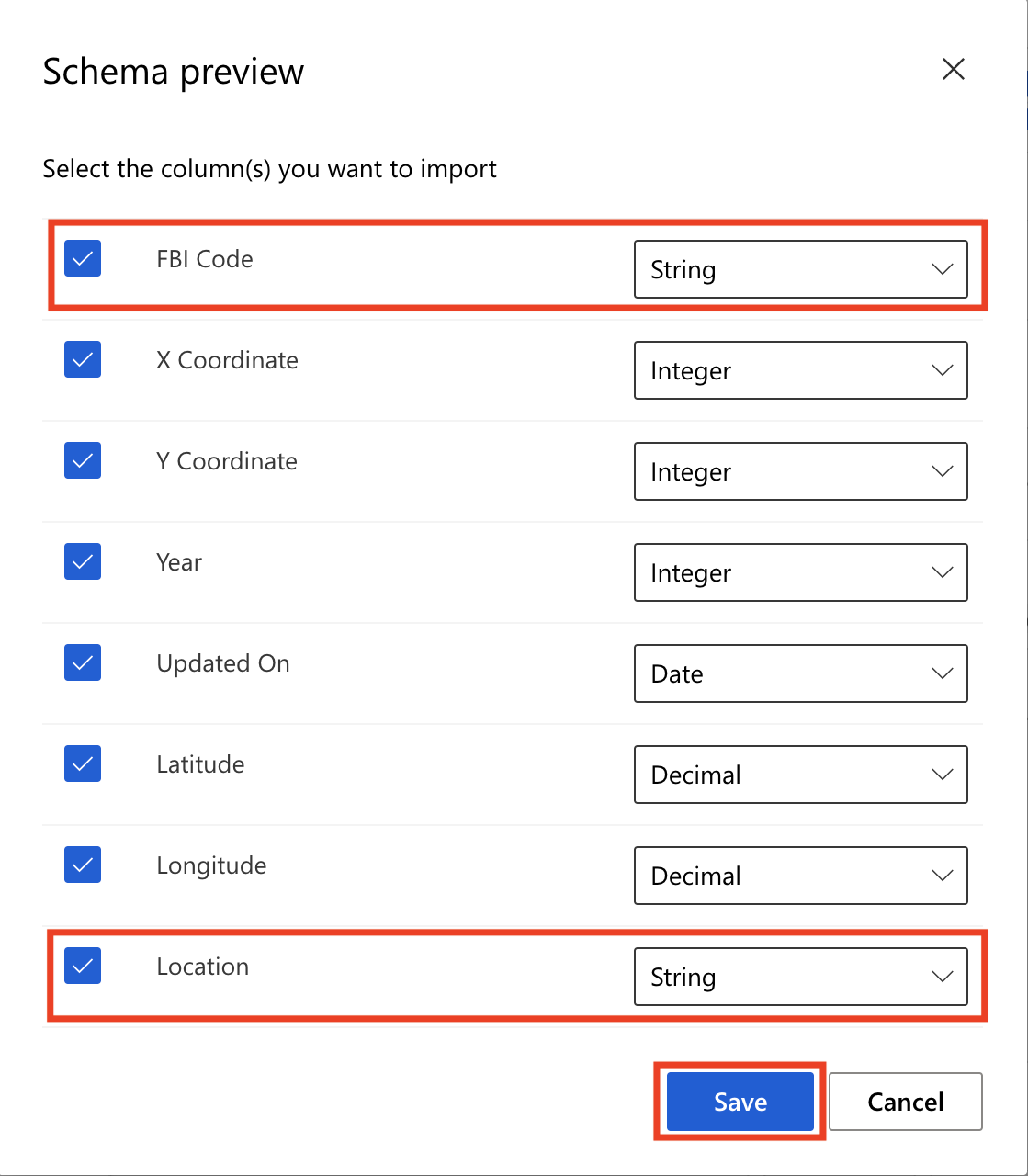
Exercise 2: Restructure the data split across multiple files

Task 1: Append rows from two additional data sources

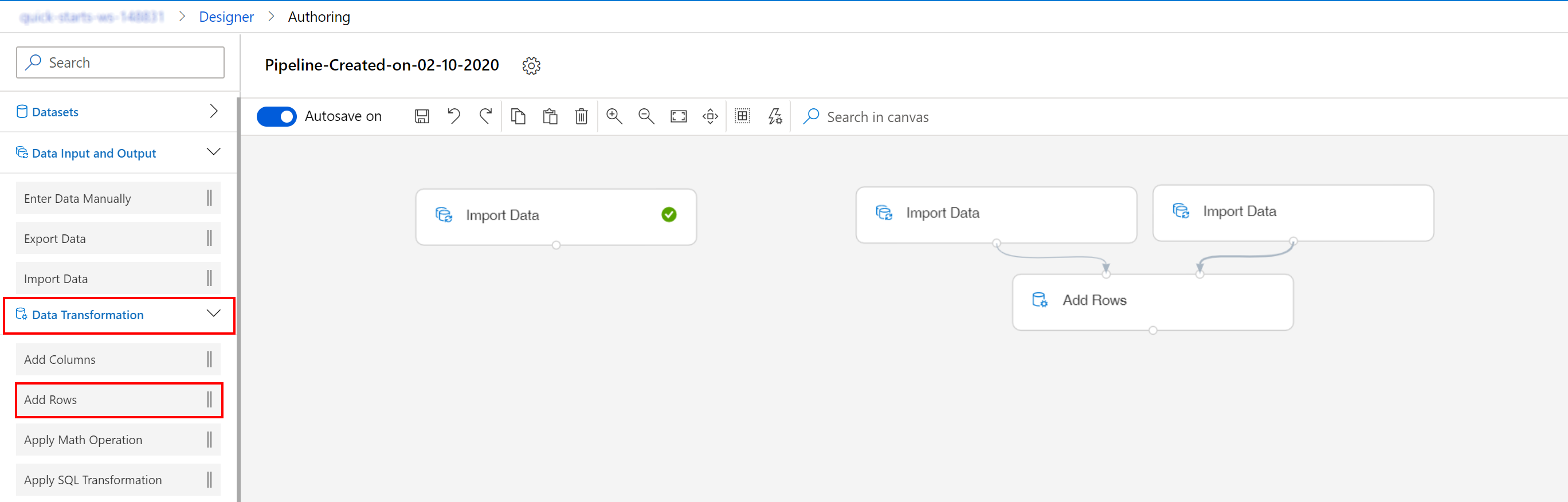
1. Select **Data Input and Output** section in the left navigation. Next, drag and drop two **Import Data** modules on to the canvas as demonstrated in the first exercise and fill in the Web URLs as follows:
   * for the first one, **Data source URL** : https://introtomlsampledata.blob.core.windows.net/data/crime-data/crime-spring.csv
   * for the second one, **Data source URL** : https://introtomlsampledata.blob.core.windows.net/data/crime-data/crime-winter.csv



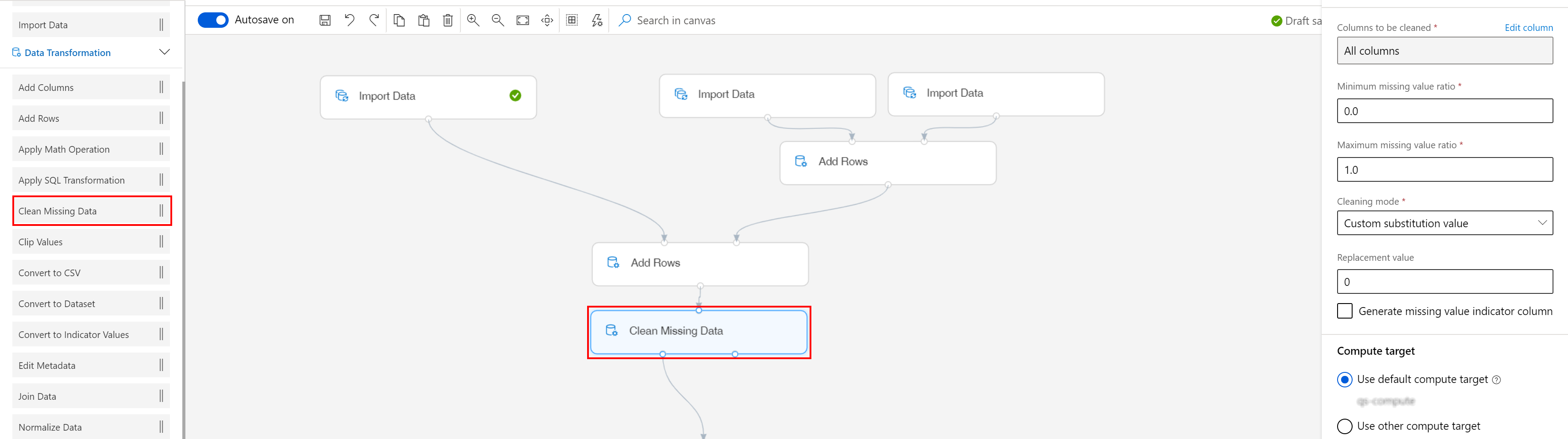
1. For each of the **three** Import Data modules, select **Preview schema** and ensure that the data type for FBI Code and Location is of type String and then select **Save**.



1. Select the **Data Transformation** section in the left navigation. Drag and drop the **Add rows** module and connect it to the above added Import data modules.

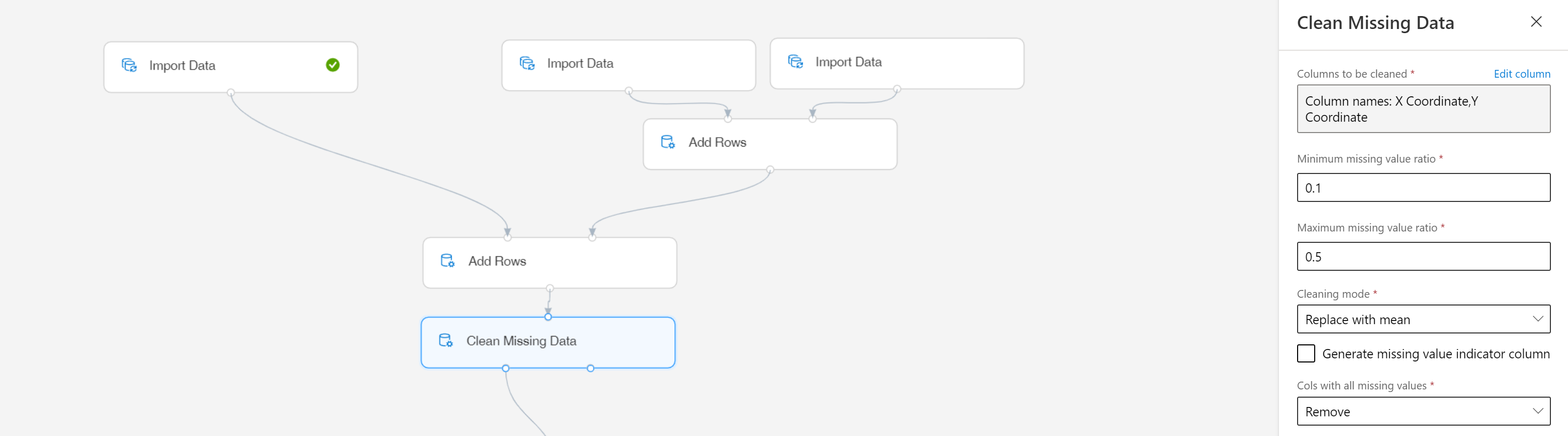


1. Repeat the same step and add a second **Add rows** module that connects the output from the first **Import data** module to the output of the first **Add rows** module.

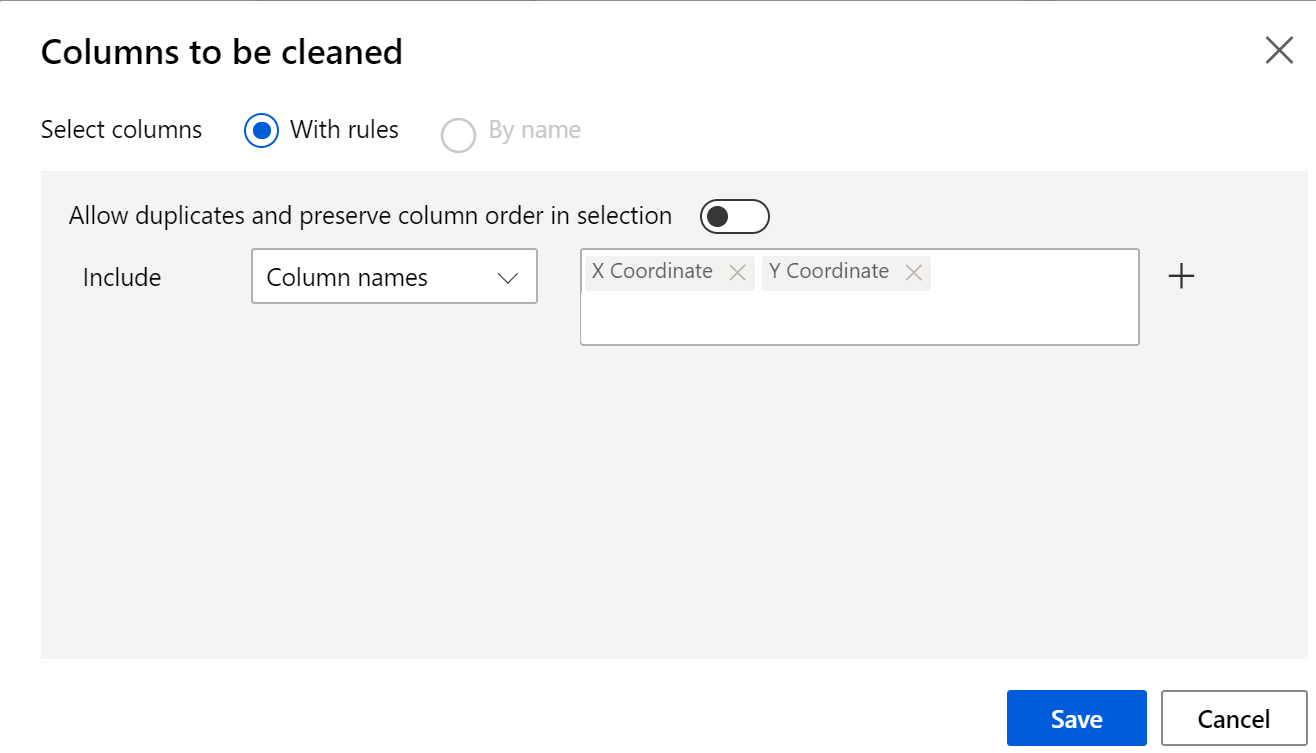


Task 2: Clean missing values

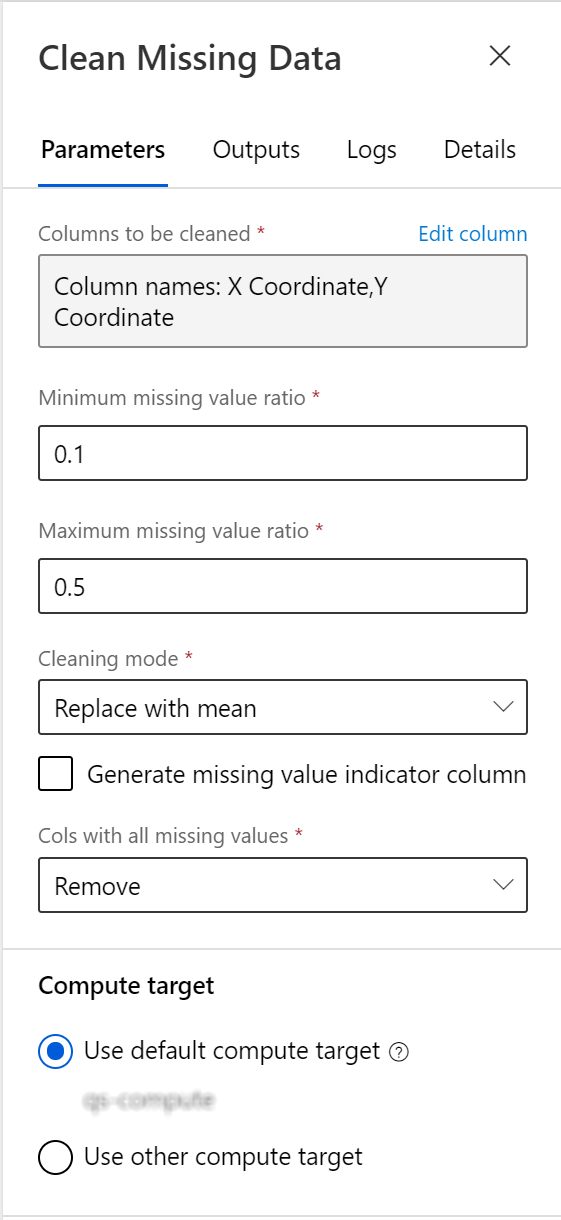
1. Drag the **Clean Missing Data** module from the **Data Transformation** section in the left navigation.



1. Select **Edit column** in the right pane to configure the list of columns to be cleaned. Select Column names from the available include options and type the name of the columns you intend to clean at this step: X Coordinate and Y Coordinate. Select **Save** to close the dialog.

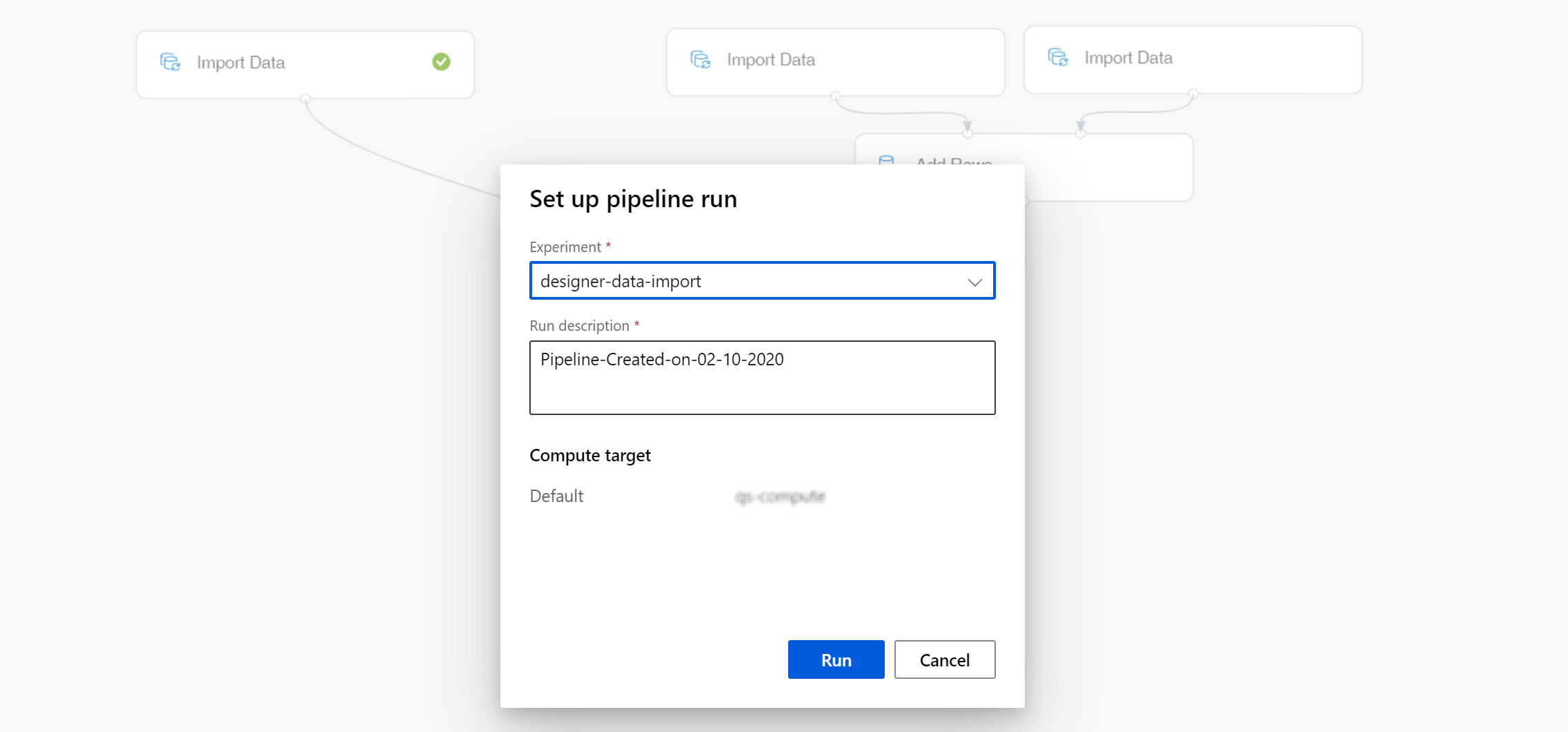


1. Set the **Minimum missing value ratio** to 0.1 and the **Maximum missing value ratio** to 0.5. Select Replace with mean in the **Cleaning mode** field.



Task 3: Submit Pipeline

1. Select **Submit** to open the Setup pipeline run editor.
2. In the Setup pipeline run editor, select **Select existing, designer-data-import** for Experiment, and then select **Submit**.

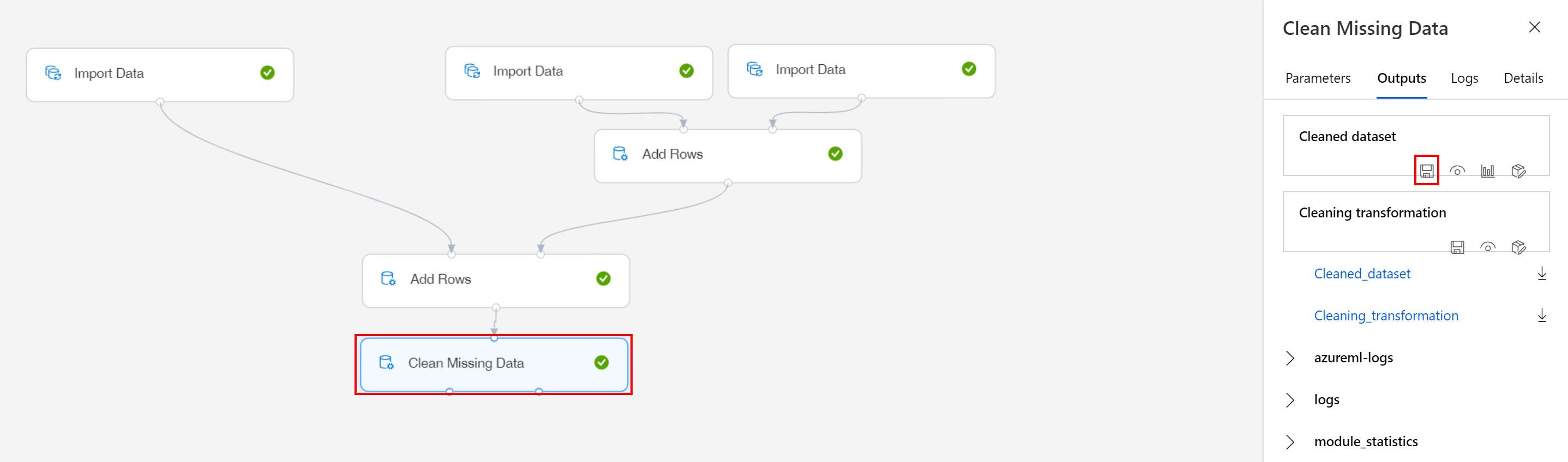


Please note that the button name in the UI is changed from **Run** to **Submit**.

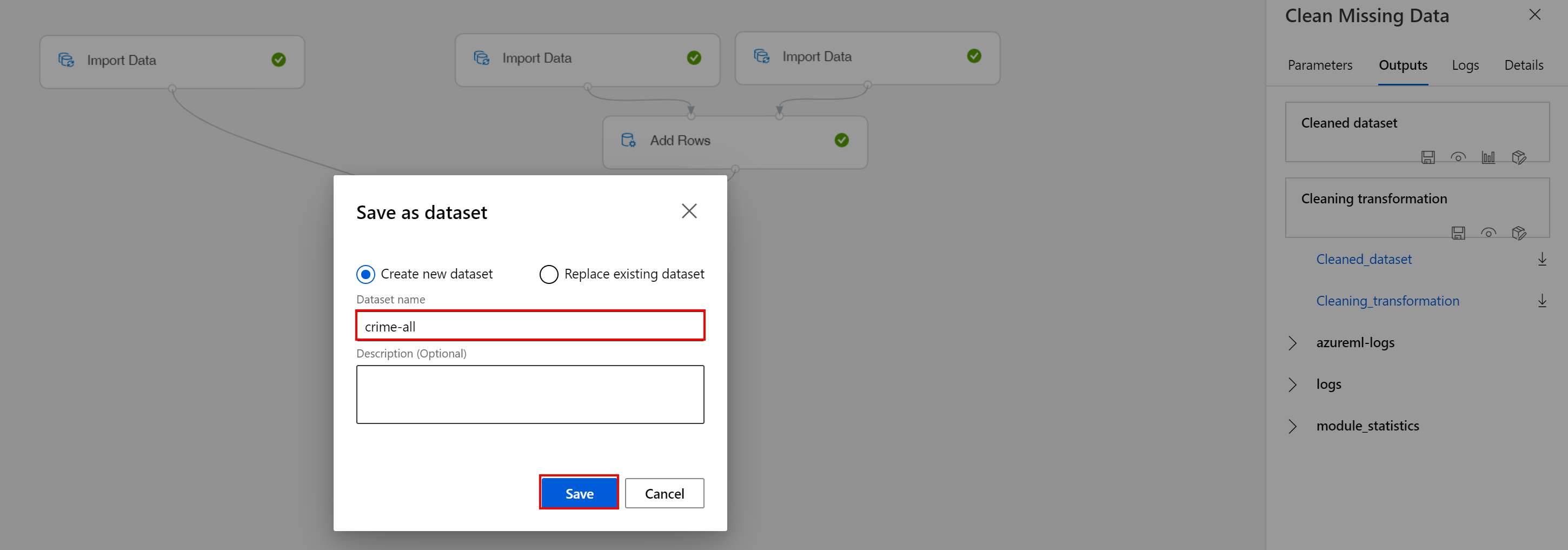
1. Wait for pipeline run to complete. It will take around **8 minutes** to complete the run.

Task 4: Save the clean dataset

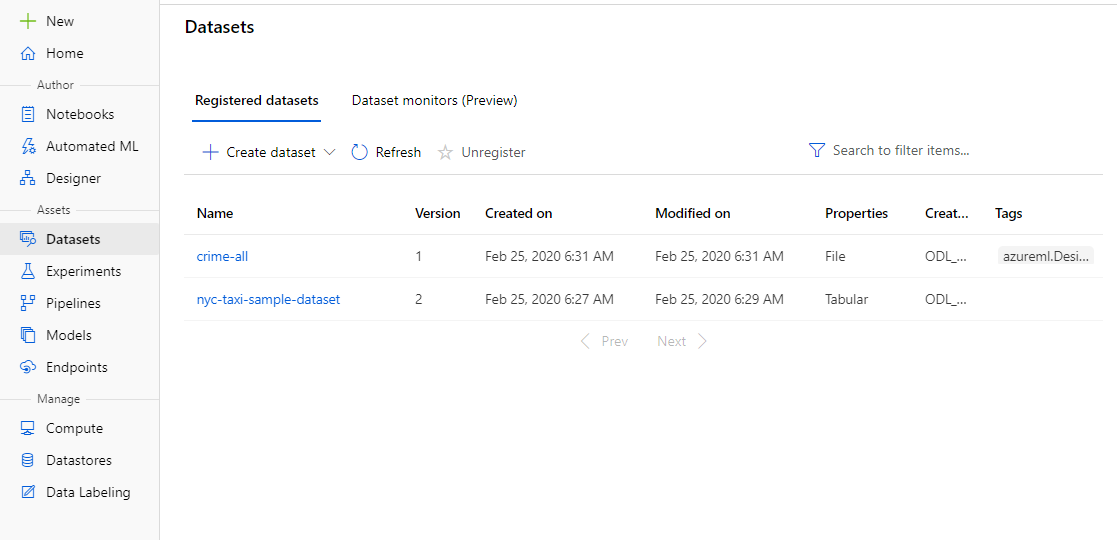
1. Select the Clean missing data module you created on the canvas and then select **Outputs + logs** on the right pane. Click on the **Save** icon under the **Cleaned dataset** section to open the Save as dataset dialog.



1. Check the option to create a new dataset and enter crime-all in the dataset name field. Select **Save** to close the dialog.



1. From the left navigation, select **Datasets**. This will open the Registered datasets page. See your registered dataset among the other datasets you used during this lesson.



Next Steps

Congratulations! You completed a few basic steps involved in the data explore and transform process, using the prebuilt modules you can find in the visual editor provided by Azure Machine Learning Studio. You can continue to experiment in the environment but are free to close the lab environment tab and return to the Udacity portal to continue with the lesson.