**Create and version a dataset**

Lab Overview

To access your data in your storage account, Azure Machine Learning offers datastores and datasets. Create an Azure Machine Learning datasets to interact with data in your datastores and package your data into a consumable object for machine learning tasks. Register the dataset to your workspace to share and reuse it across different experiments without data ingestion complexities.

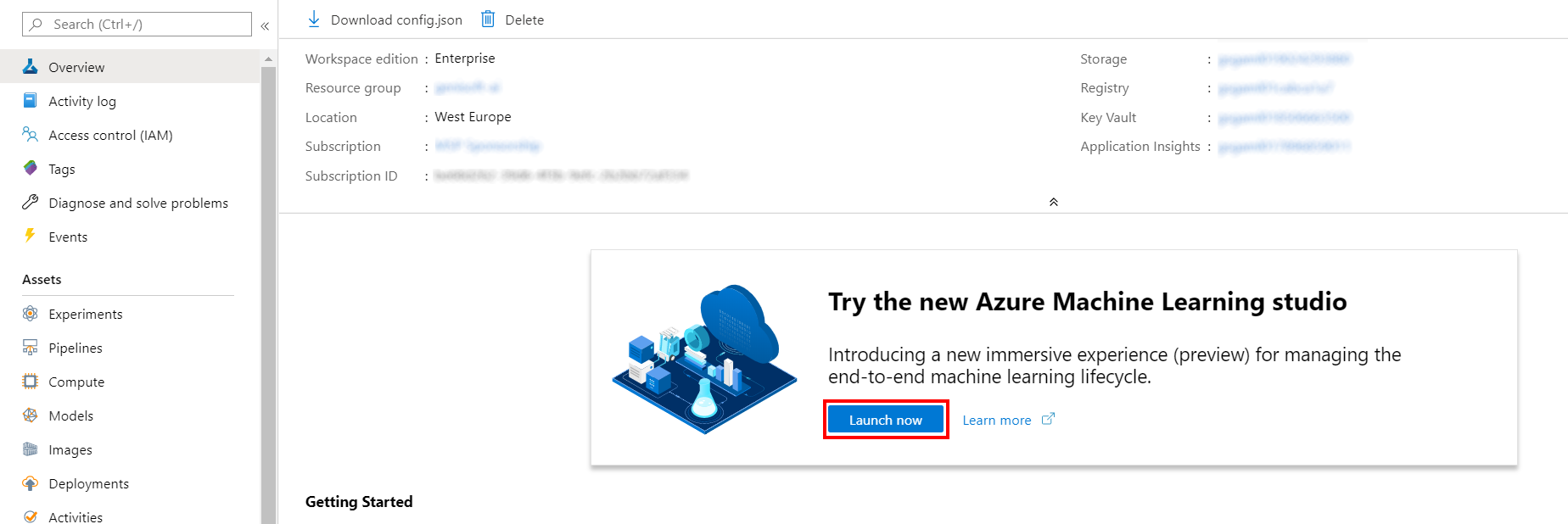
Datasets can be created from local files, public urls, Azure Open Datasets, or specific file(s) in your datastores. To create a dataset from an in memory pandas dataframe, write the data to a local file, like a csv, and create your dataset from that file. Datasets aren’t copies of your data, but are references that point to the data in your storage service, so no extra storage cost is incurred.

In this lab, we are using a subset of NYC Taxi & Limousine Commission - green taxi trip records available from [Azure Open Datasets](https://azure.microsoft.com/en-us/services/open-datasets/) to show how you can register and version a Dataset using the AML designer interface. In the first exercises we use a modified version of the original CSV file, which includes collected records for five months (January till May). The second exercise demonstrates how we can create a new version of the initial dataset when new data is collected (in this case, we included records collected in June in the CSV file).

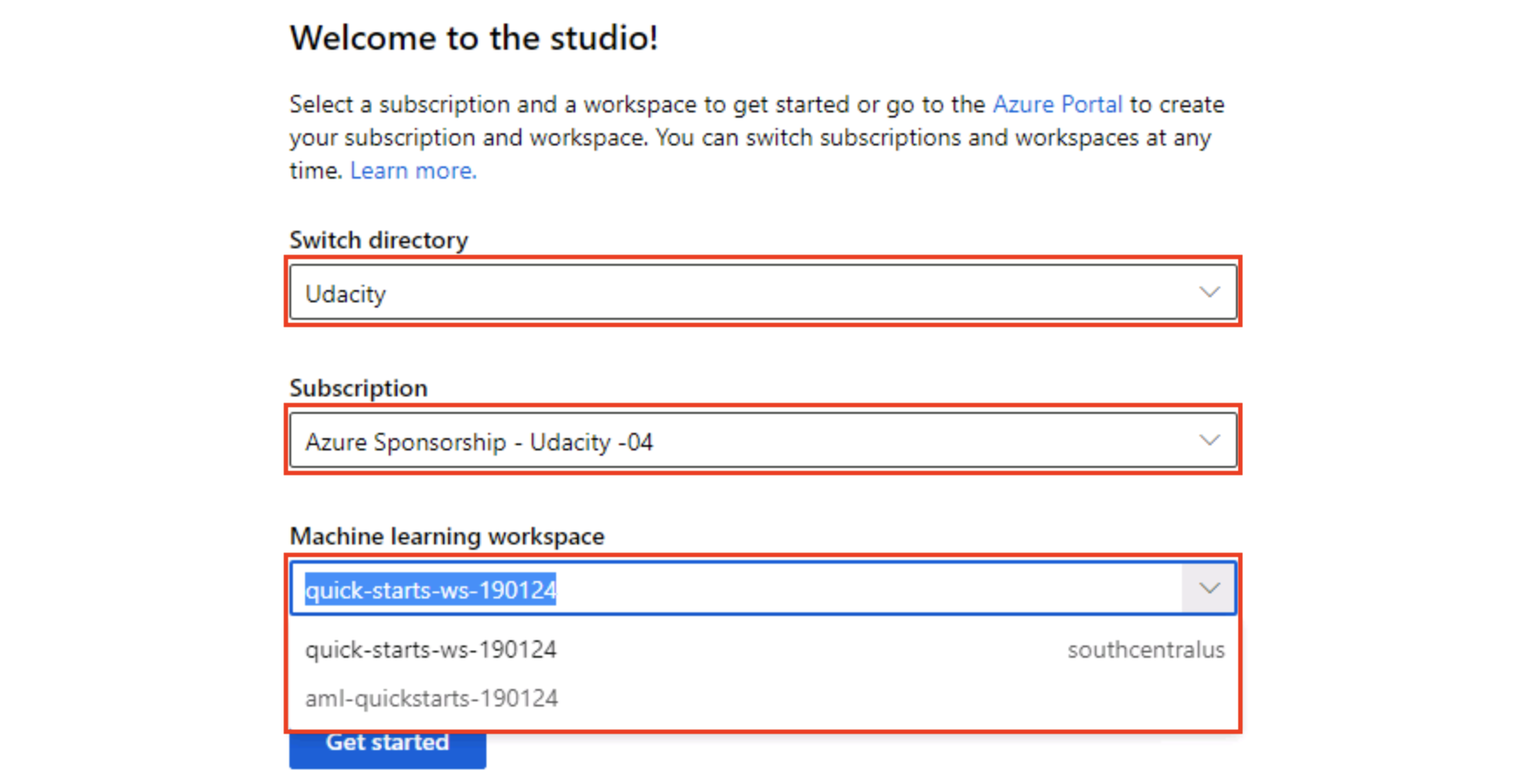
Exercise 1: Register Dataset with Azure Machine Learning studio

Task 1: Upload Dataset from web file

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.

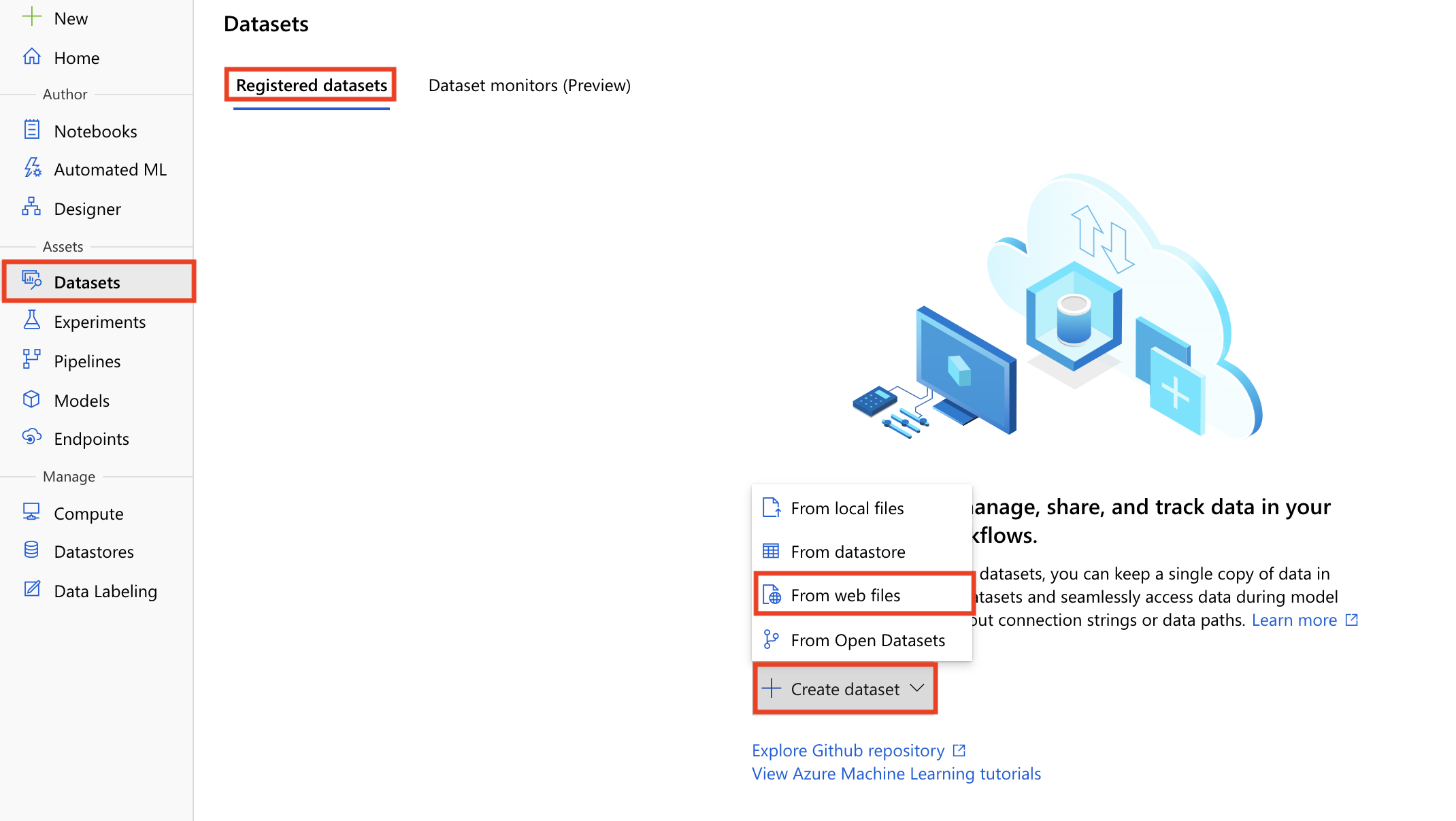


1. When you first launch the studio, you may need to set the directory and subscription. If so, you will see this screen:

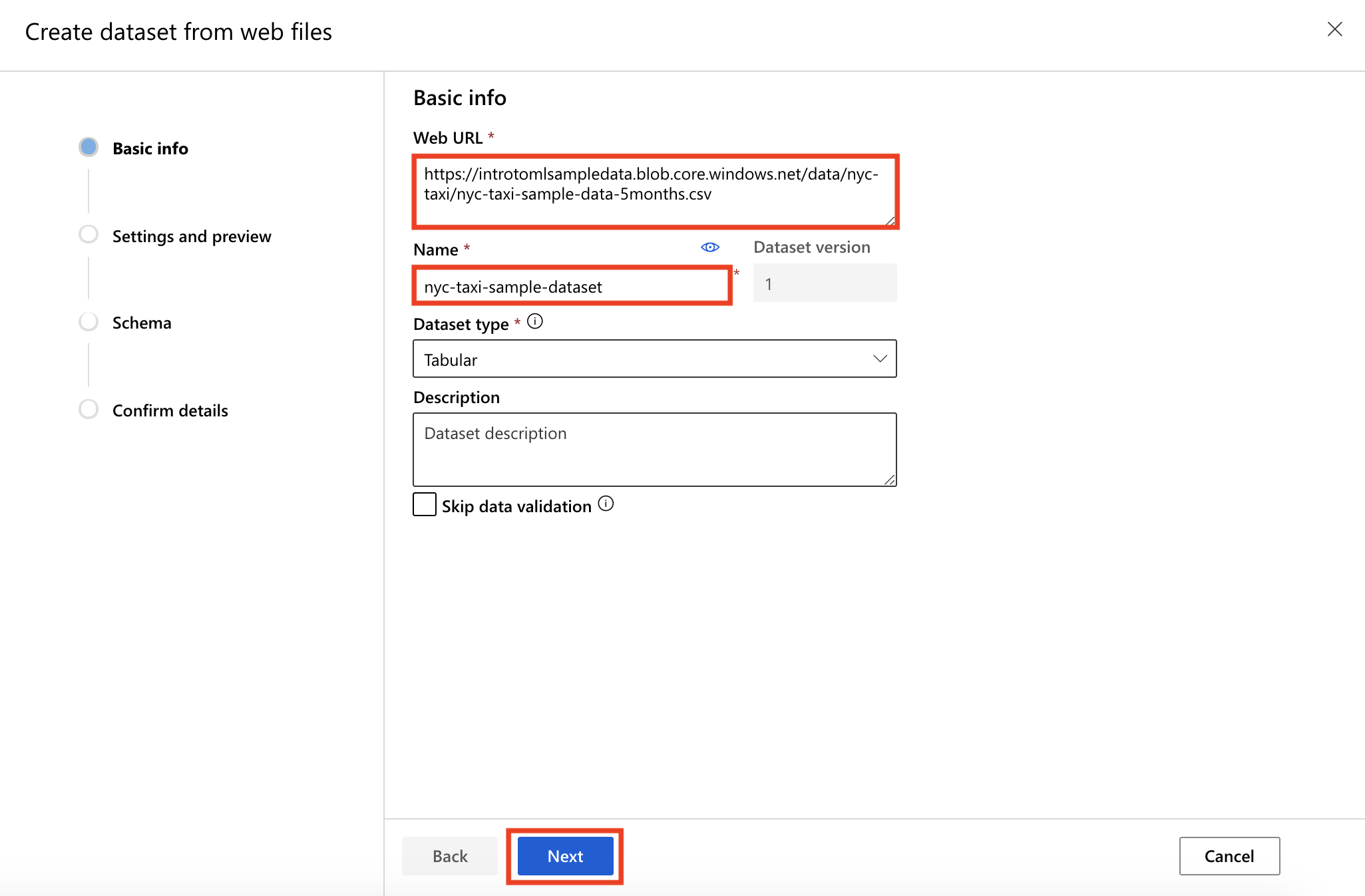


For the directory, select **Udacity** and for the subscription, select **Azure Sponsorship**. For the machine learning workspace, you may see multiple options listed. **Select any of these** (it doesn’t matter which) and then click **Get started**.

1. From the studio, select **Datasets, + Create dataset, From web files**. This will open the Create dataset from web files dialog on the right.

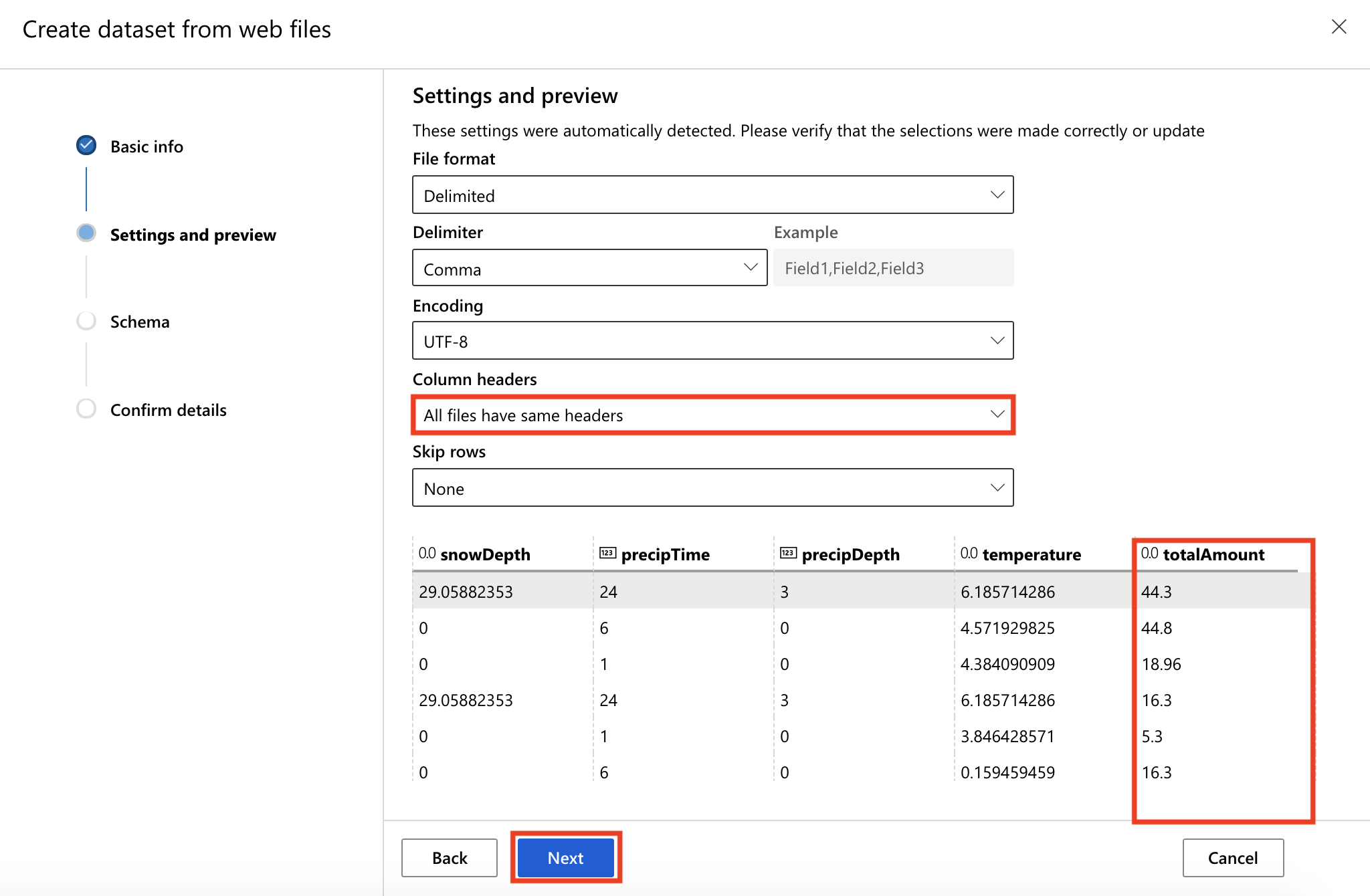


1. Provide the following information and then select **Next**:
   1. Web URL: https://introtomlsampledata.blob.core.windows.net/data/nyc-taxi/nyc-taxi-sample-data-5months.csv
   2. Name: nyc-taxi-sample-dataset



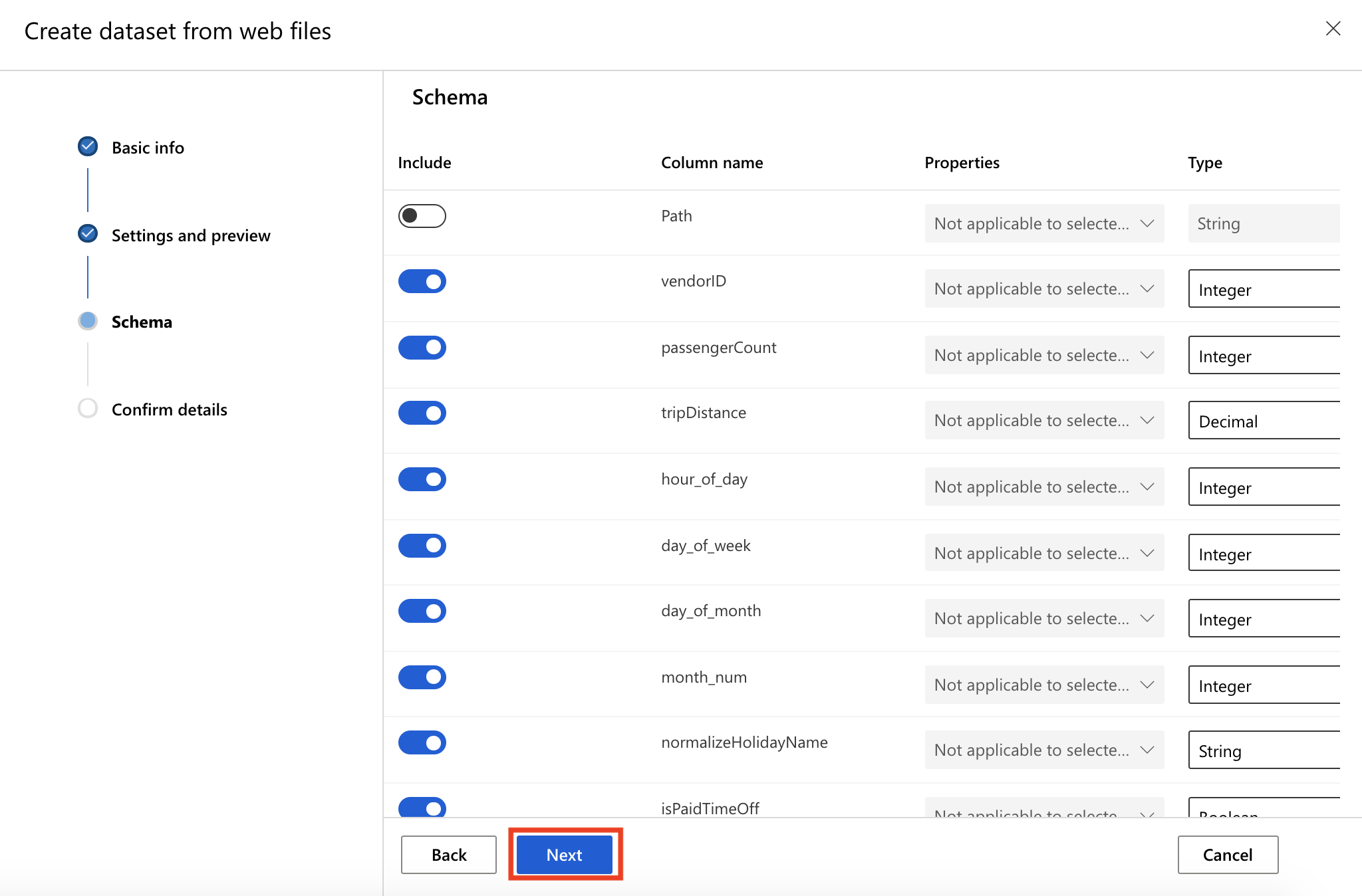
Task 2: Preview Dataset

1. On the Settings and preview panel, set the **Column headers** drop down to All files have same headers.
2. Scroll the data preview to right to observe the target column: totalAmount. After you are done reviewing the data, select **Next**



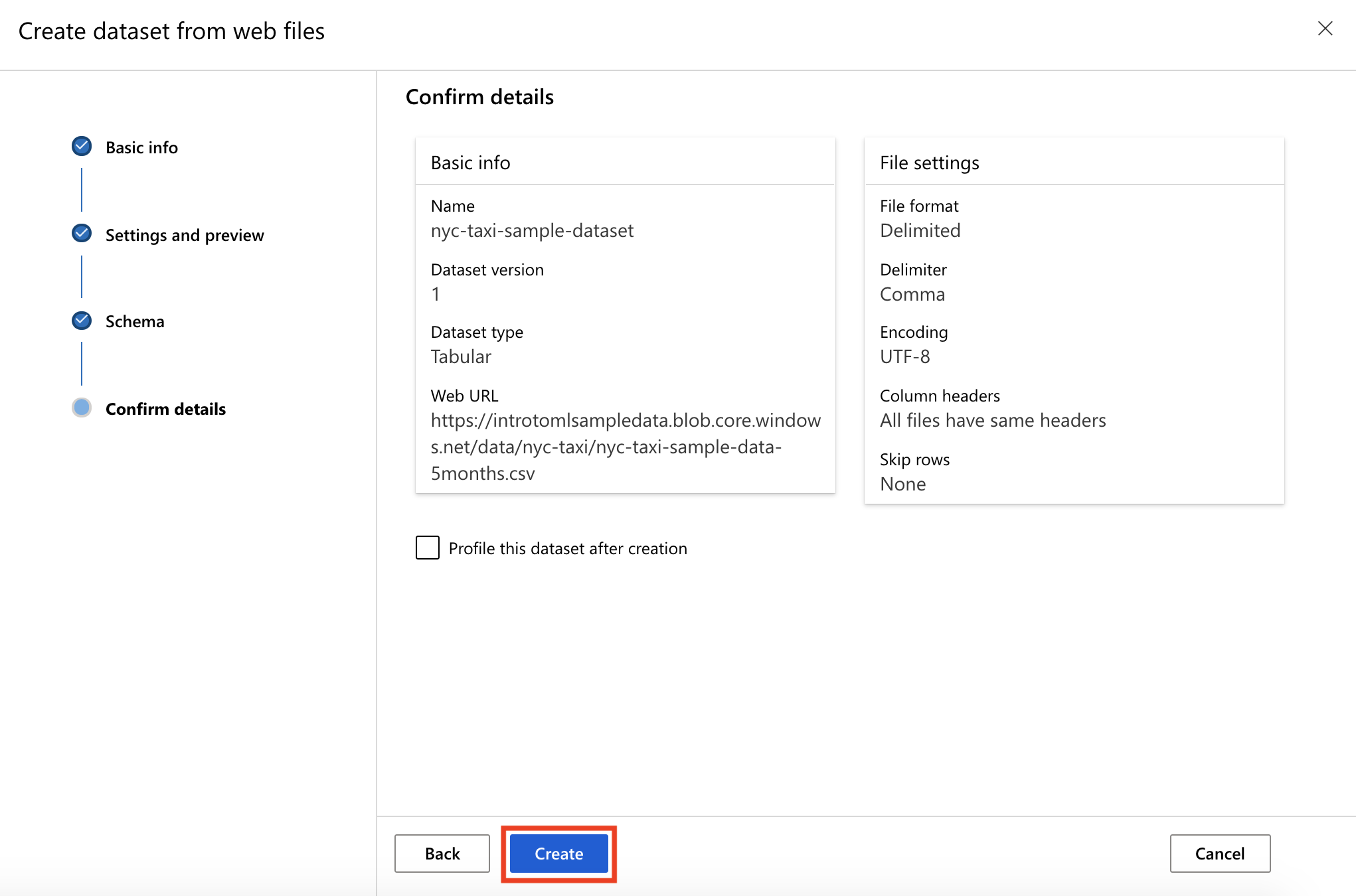
Task 3: Select Columns

1. Select columns from the dataset to include as part of your training data. Leave the default selections and select **Next**



Task 4: Create Dataset

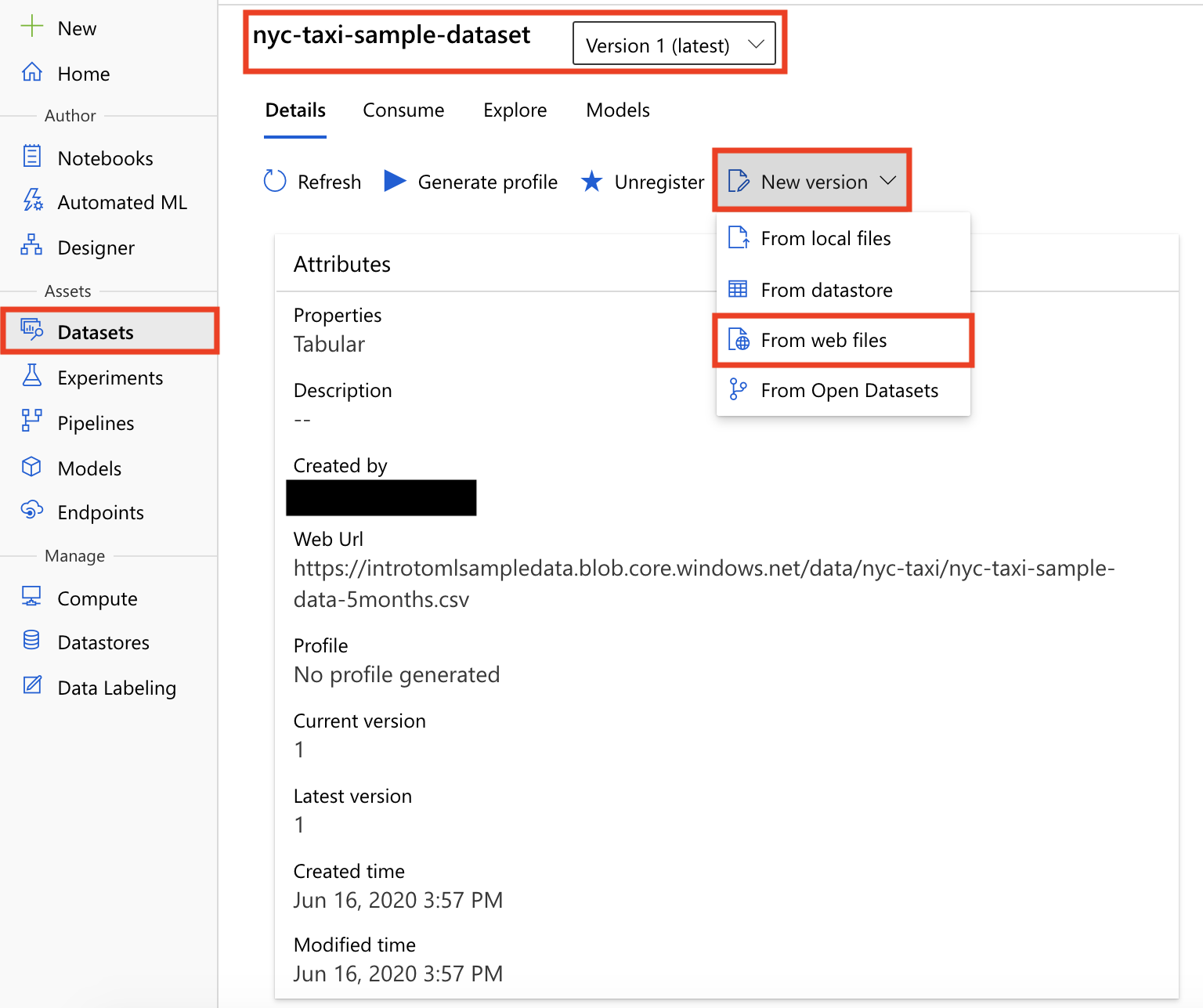
1. Confirm the dataset details and select **Create**



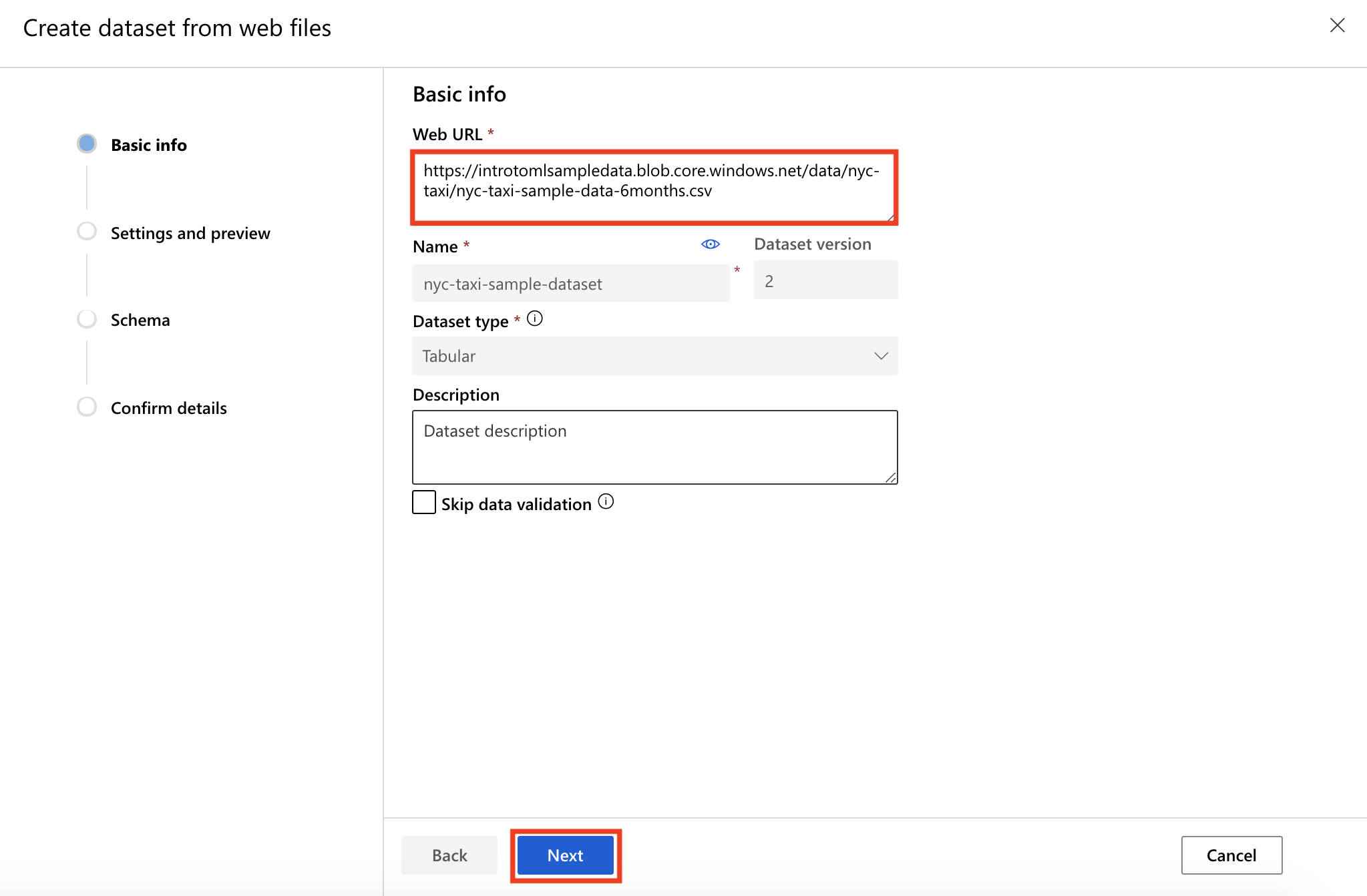
Exercise 2: Create a version of the existing Dataset

Task 1: Register new dataset version

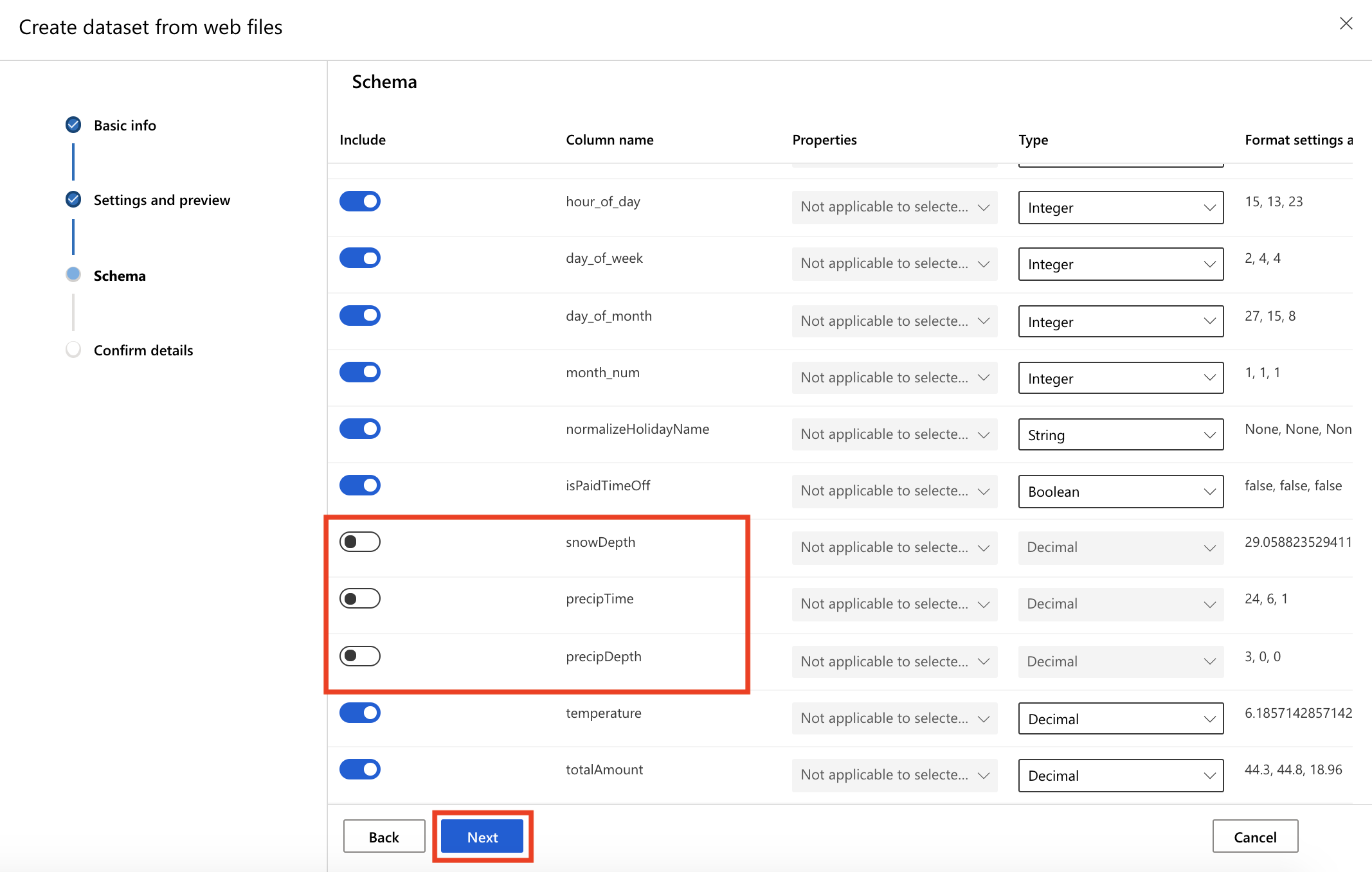
1. From the [Azure Machine Learning studio](https://ml.azure.com/), select **Datasets** and select the nyc-taxi-sample-dataset dataset created in the first exercise. This will open the Dataset details page.
2. Select **New version, From web files** to open the same Create dataset from web files dialog you already entered in the first exercise.



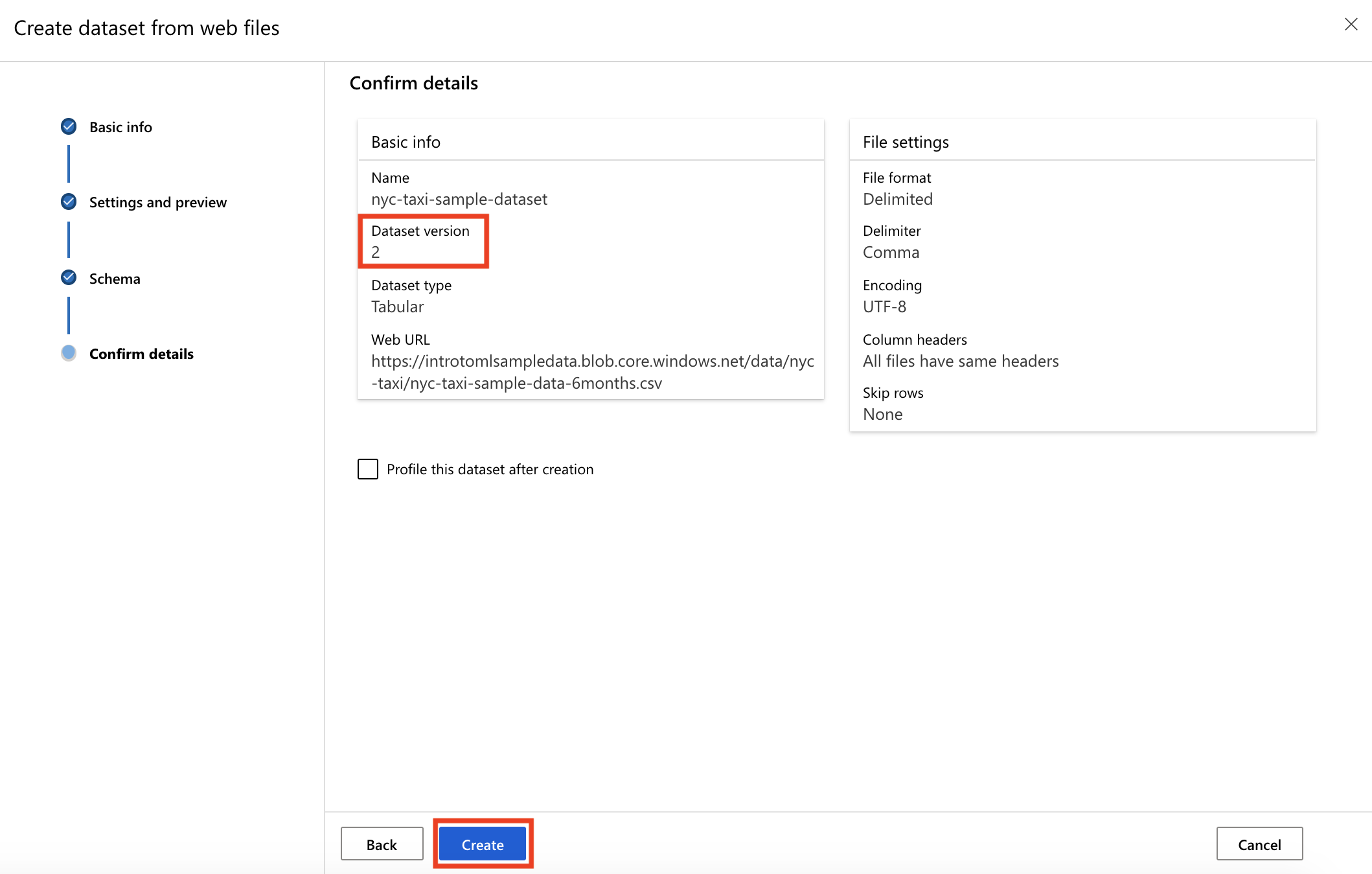
1. This time, the **Name** and **Dataset version** fields are already filled in for you. Provide the following information and select **Next** to move on to the next step:
   1. Web URL: https://introtomlsampledata.blob.core.windows.net/data/nyc-taxi/nyc-taxi-sample-data-6months.csv



1. Select All files have the same headers in the **Column headers** drop-down and move on to the schema selection step.
2. On the Schema page, let’s suppose you decided to exclude some columns from your dataset. Exclude columns: **snowDepth**, **prcipTime**, **precipDepth**. Select **Next** to move on to the final step.

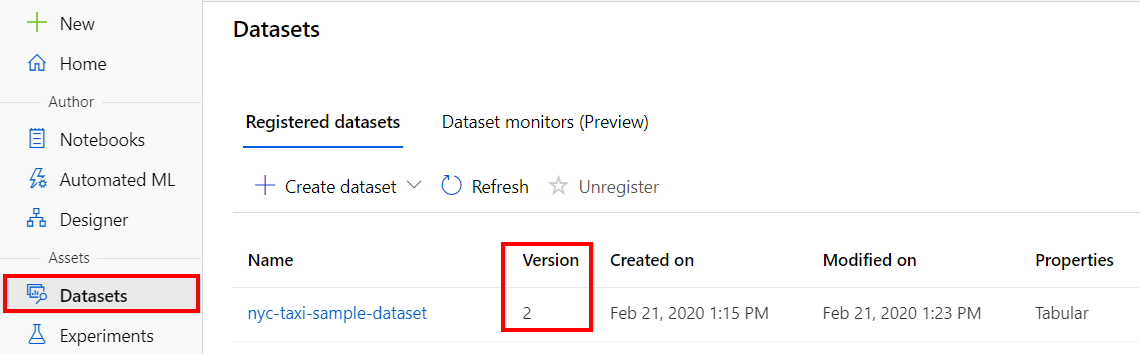


1. Notice the Dataset version value in the basic info section. Select **Create** to close the new version confirmation page.

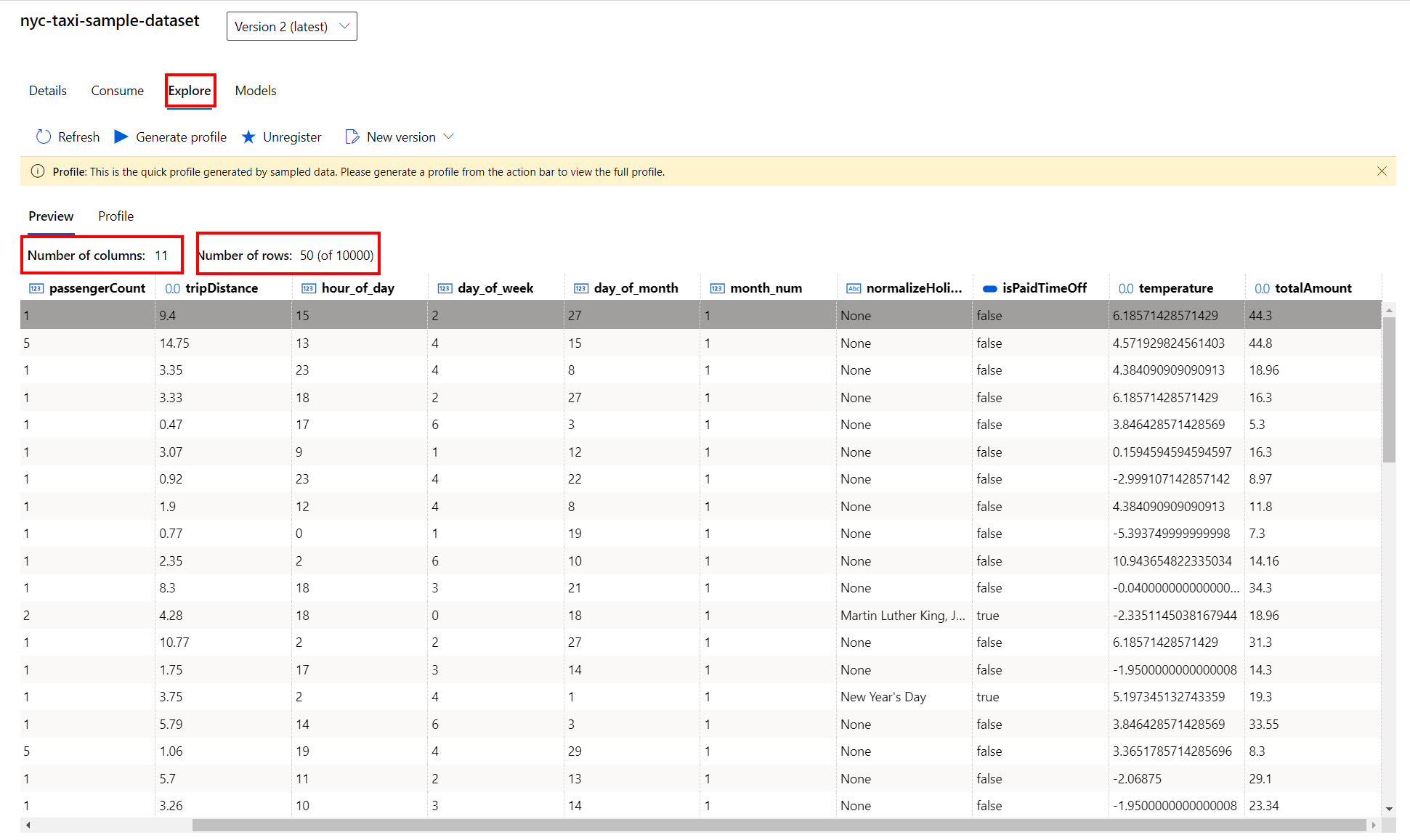


Task 2: Review both versions of the dataset

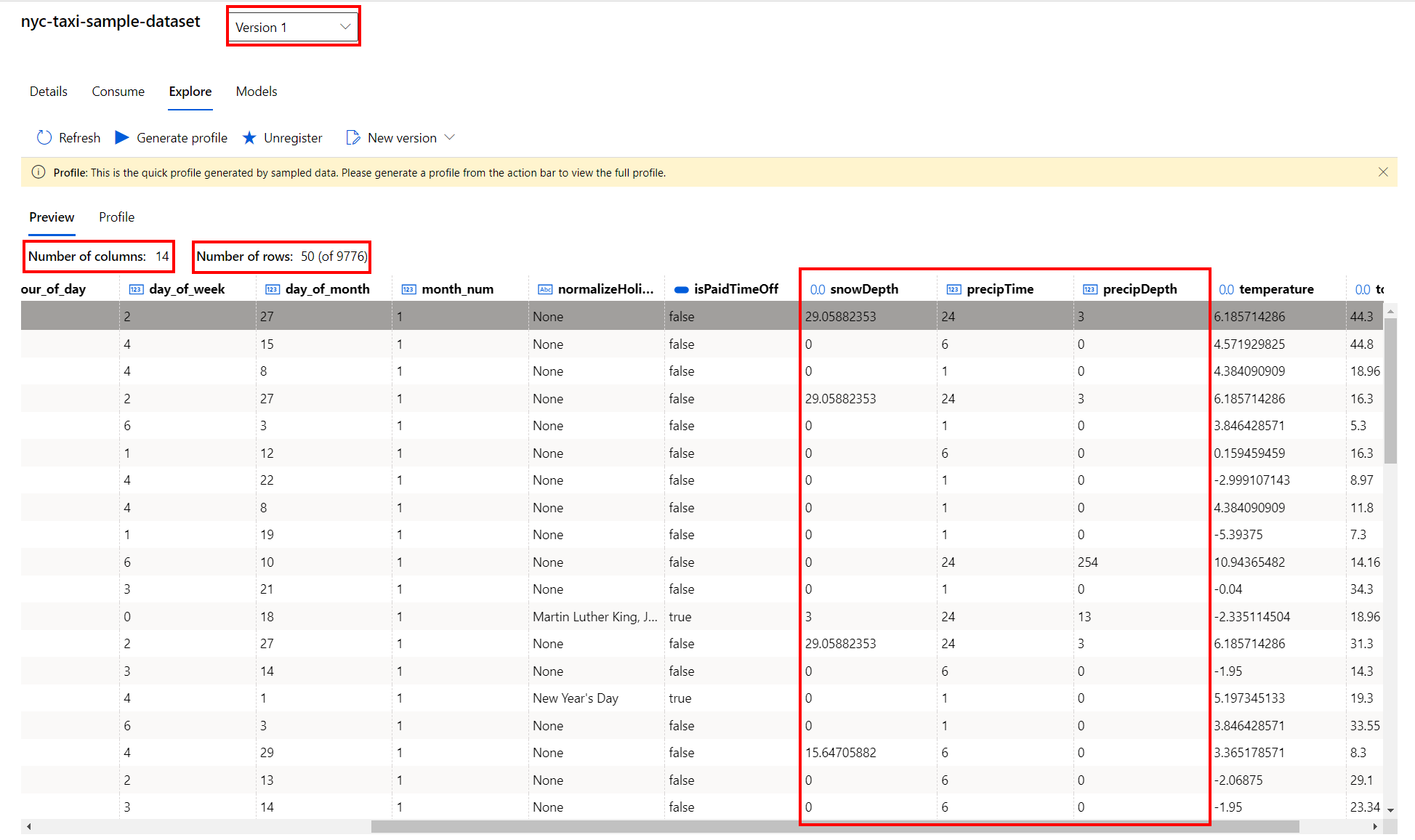
1. Back to the **Datasets** page, in the **Registered datasets** list, notice the version value for the nyc-taxi-sample-dataset dataset.



1. Select the nyc-taxi-sample-dataset dataset link to open the dataset details page, where **Version 2(latest)** is automatically selected. Go to the **Explore** section to observe the structure and content of the new version. Notice the columns and rows structure in the dataset preview pane:
   * **Number of columns**: 11
   * **Number of rows**: 10000
   * Scroll right to check that the three excluded columns are missing (**snowDepth**, **prcipTime**, **precipDepth**)



1. Select **Version 1** from the drop-down near the dataset name title and notice the changing values for:
   * **Number of columns**: 14 (since the previous version still contains the three excluded columns)
   * **Number of rows**: 9776 (since the previous version contains only data for 5 months)



Next Steps

Congratulations! You have now explored a first simple scenario for dataset versioning using the Azure Machine Learning studio. You found out how you can create and version a simple dataset when new training data is available. 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.