**Prepare and visualize data without writing code in Azure Machine Learning**

# Introduction :

Prepare and visualize your data in the drag-and-drop visual interface (preview) for Azure Machine Learning. The data you'll use includes entries for various individual automobiles, including information such as make, model, technical specifications, and price.

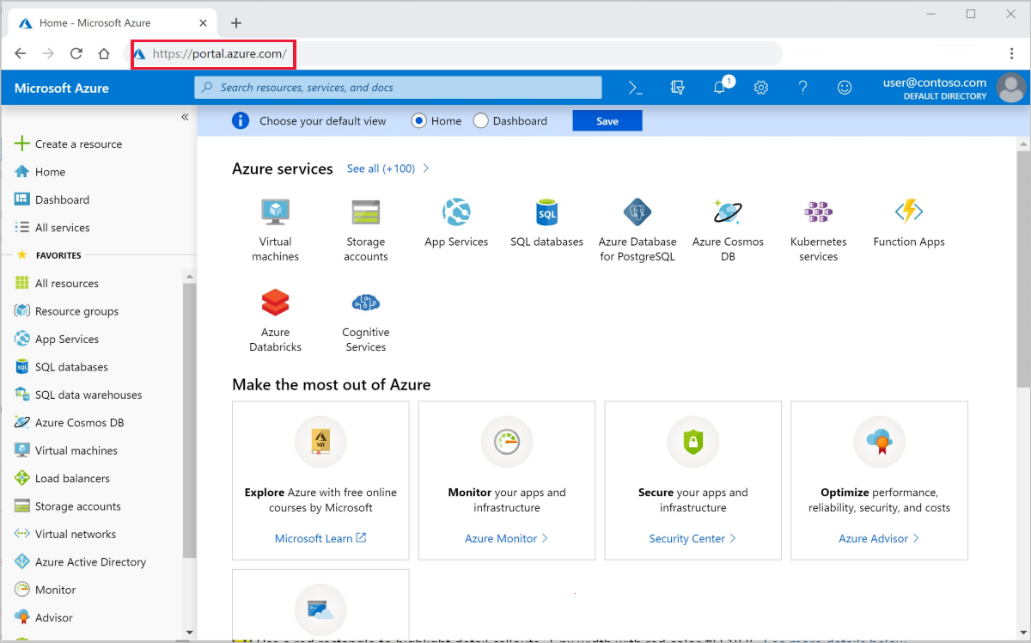
In this quickstart you'll explore and prepare data:

* Create your first experiment to add and preview data
* Prepare the data by removing missing values
* Run the experiment
* Visualize the resulting data

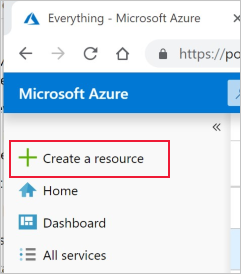
If you're brand new to machine learning, the video series [Data Science for Beginners](https://docs.microsoft.com/azure/machine-learning/studio/data-science-for-beginners-the-5-questions-data-science-answers) is a great introduction to machine learning.

# Create a workspace

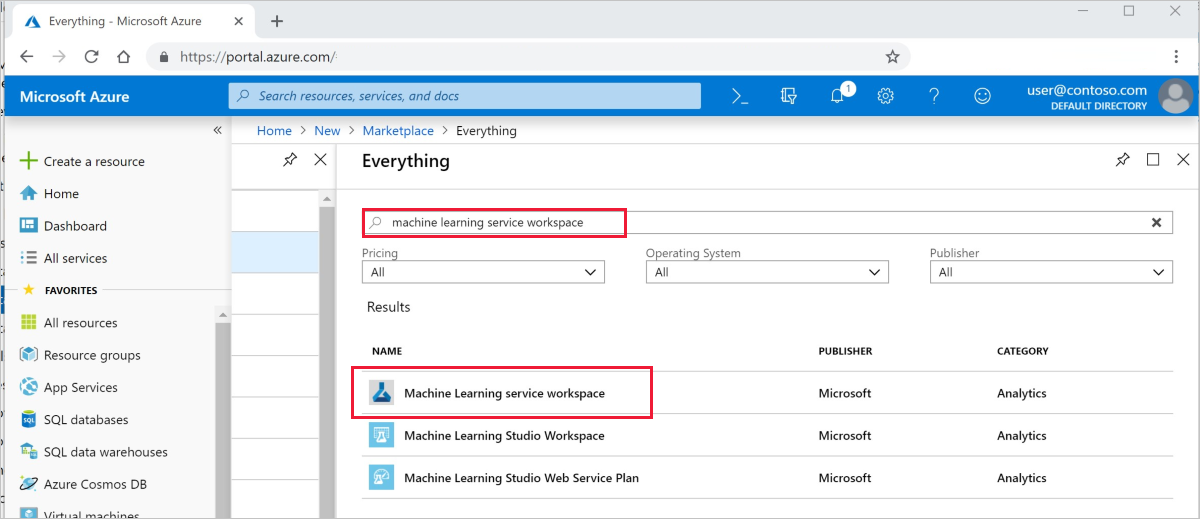
1. Sign in to the [Azure portal](https://portal.azure.com/) by using the credentials for the Azure subscription you use.



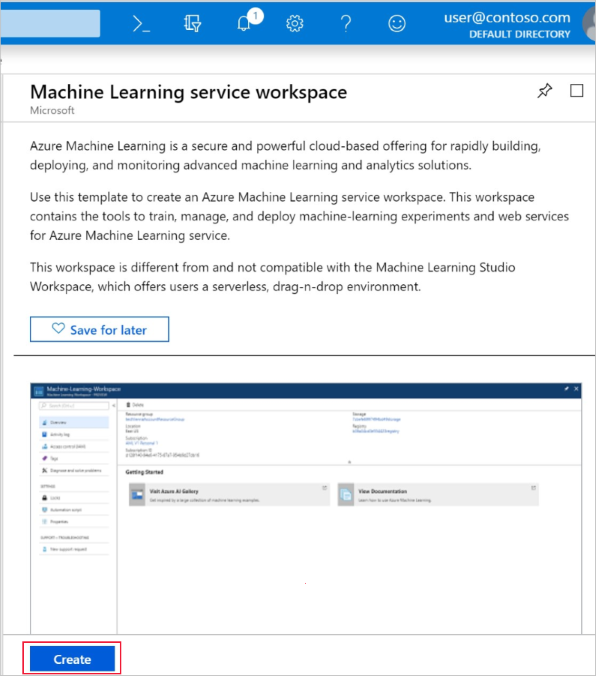
1. In the upper-left corner of the portal, select **Create a resource**.



1. In the search bar, enter **Machine Learning**. Select the **Machine Learning service workspace** search result.

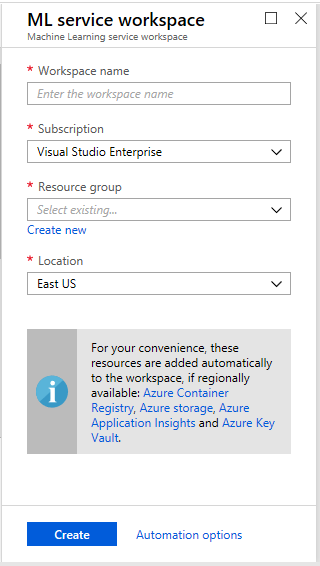


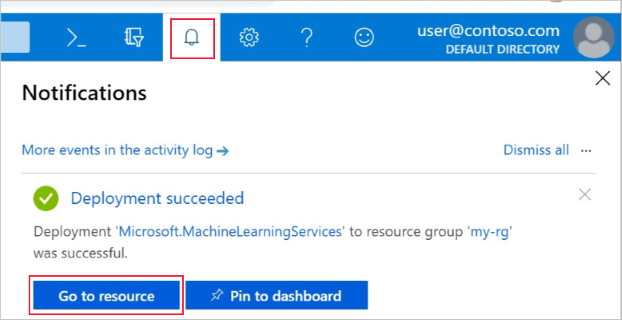
1. In the **ML service workspace** pane, scroll to the bottom and select **Create** to begin.



1. In the **ML service workspace** pane, configure your workspace.

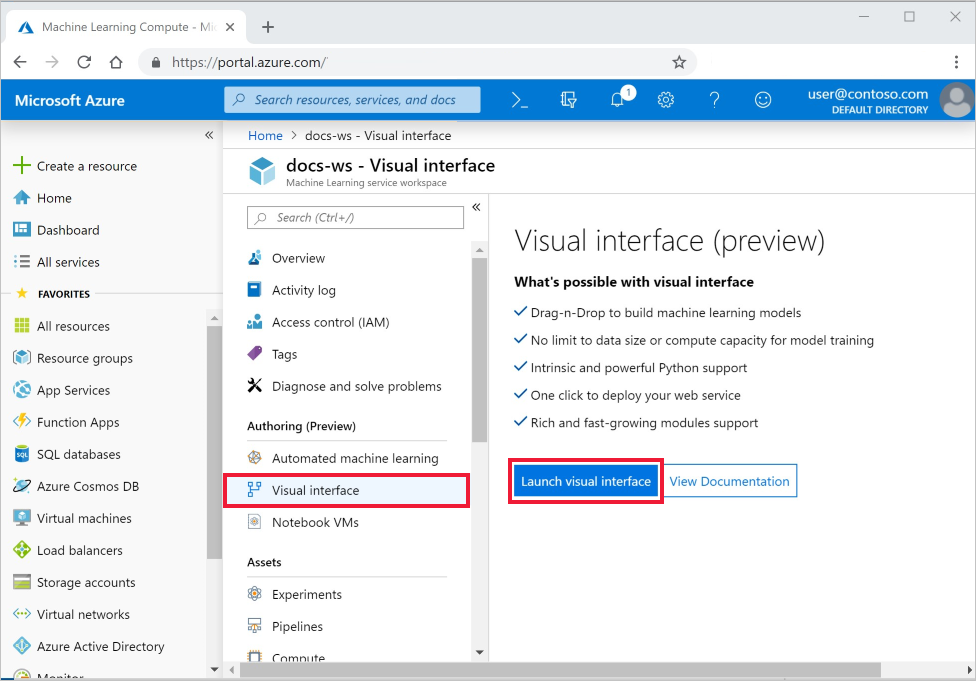
| **Field** | **Description** |
| --- | --- |
| Workspace name | Enter a unique name that identifies your workspace. In this example, we use **docs-ws**. Names must be unique across the resource group. Use a name that's easy to recall and differentiate from workspaces created by others. |
| Subscription | Select the Azure subscription that you want to use. |
| Resource group | Use an existing resource group in your subscription, or enter a name to create a new resource group. A resource group is a container that holds related resources for an Azure solution. In this example, we use **docs-aml**. |
| Location | Select the location closest to your users and the data resources. This location is where the workspace is created. |

1. 
2. To start the creation process, select **Create**. It can take a few moments to create the workspace.
3. To check on the status of the deployment, select the Notifications icon, **bell**, on the toolbar.
4. When the process is finished, a deployment success message appears. It's also present in the notifications section. To view the new workspace, select **Go to resource**.



## Open the visual interface webpage

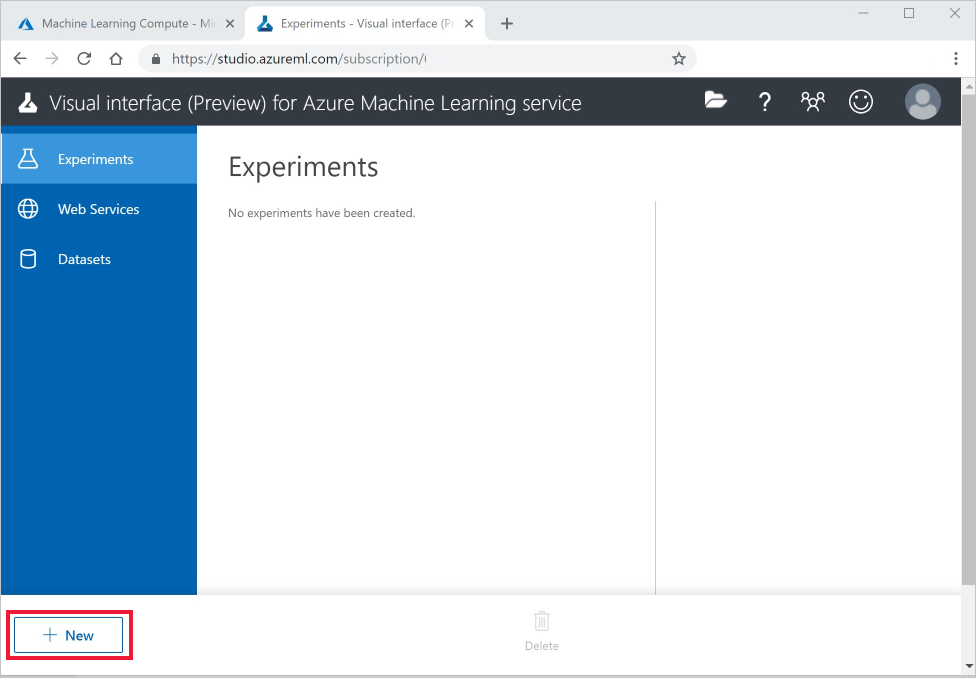
1. Open your workspace in the [Azure portal](https://portal.azure.com/).
2. In your workspace, select **Visual interface**. Then select **Launch visual interface**.

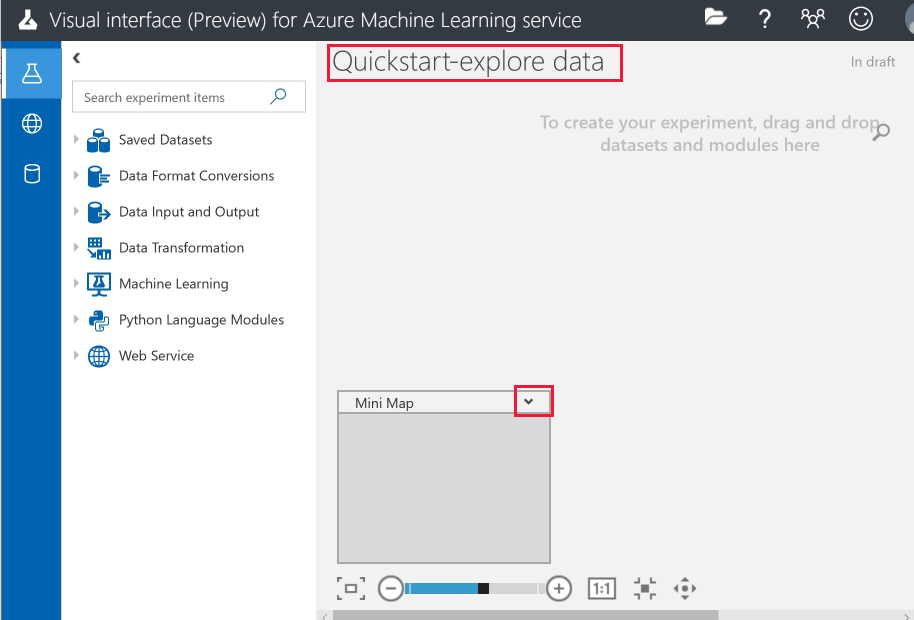


The interface webpage opens in a new browser page.

# Create your first experiment

The visual interface tool provides an interactive, visual place to easily build, test, and iterate on a predictive analysis model. You drag-and-drop datasets and analysis modules onto an interactive canvas, connecting them together to form an experiment. Create your first experiment now.

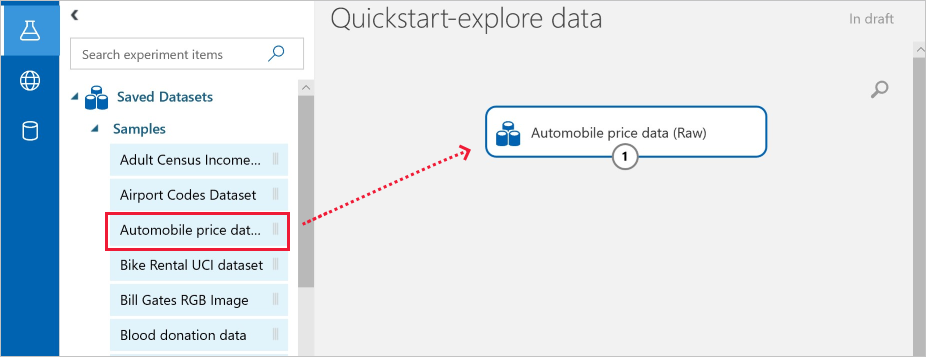
1. In the bottom-left corner, select **Add New**. 
2. Select **Blank Experiment**.
3. Your experiment is given a default name. Select this text and rename it to "Quickstart-explore data." This name doesn't need to be unique.
4. The **Mini Map** at the bottom of the screen is useful for viewing large experiments. You won't need it in this quickstart so click on the arrow at the top to minimize it.



## Add data

The first thing you need for machine learning is data. There are several sample datasets included in this interface that you can use, or you can import data from many sources. For this example, you'll use the sample dataset **Automobile price data (Raw)**.

1. To the left of the experiment canvas is a palette of datasets and modules. Select **Saved Datasets** then select **Samples** to view the available sample datasets.
2. Select the dataset, **Automobile price data (raw)**, and drag it onto the canvas.



# Select columns

Select which columns of data to work with. To start with, configure the module to show all available columns.

**Tip**

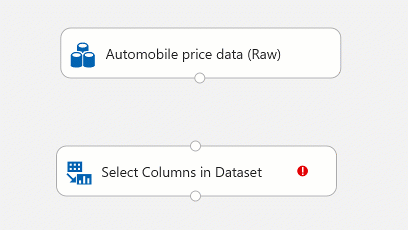
If you know the name of the data or module you want, use the search bar at the top of the palette to find it quickly. The rest of the quickstart will use this shortcut.

1. Type **Select** in the Search box to find the **Select Columns in Dataset** module.
2. Click and drag the **Select Columns in Dataset** onto the canvas. Drop the module below the dataset you added earlier.
3. Connect the dataset to the **Select Columns in Dataset**: click the output port of the dataset, drag to the input port of **Select Columns in Dataset**, then release the mouse button. The dataset and module remain connected even if you move either around on the canvas.

**Tip**

Datasets and modules have input and output ports represented by small circles - input ports at the top, output ports at the bottom. You create a flow of data through your experiment when you connect the output port of one module to an input port of another.

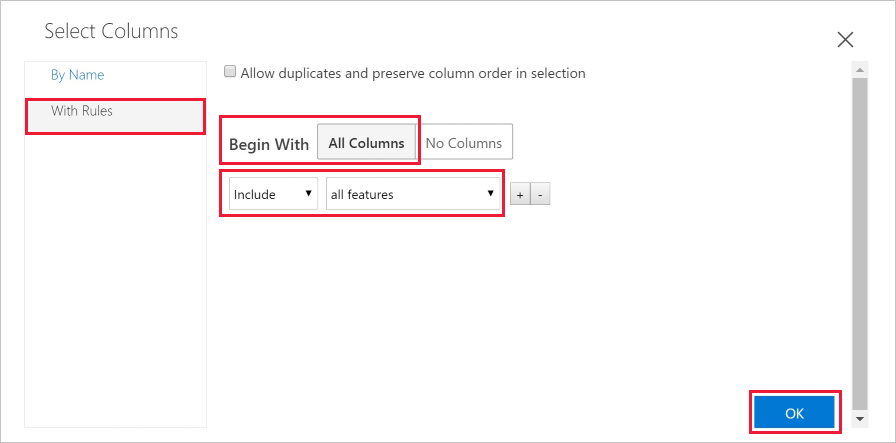
If you have trouble connecting modules, try dragging all the way into the node you are connecting.



The red exclamation mark indicates that you haven't set the properties for the module yet. You'll do that next.

1. Select the **Select Columns in Dataset** module.
2. In the **Properties** pane to the right of the canvas, select **Edit columns**.

In the **Select columns** dialog, select **ALL COLUMNS** and include **all features**. The dialog should look like this:



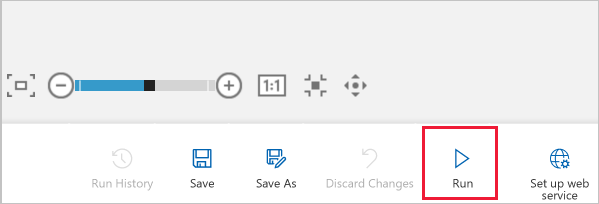
1. On the lower right, select **OK** to close the column selector.

## Run the experiment

At any time, click the output port of a dataset or module to see what the data looks like at that point in the data flow. If the **Visualize** option is disabled, you first need to run the experiment. You'll do that next.

An experiment runs on a compute target, a compute resource that is attached to your workspace. Once you create a compute target, you can reuse it for future runs.

1. Select **Run** at the bottom to run the experiment.

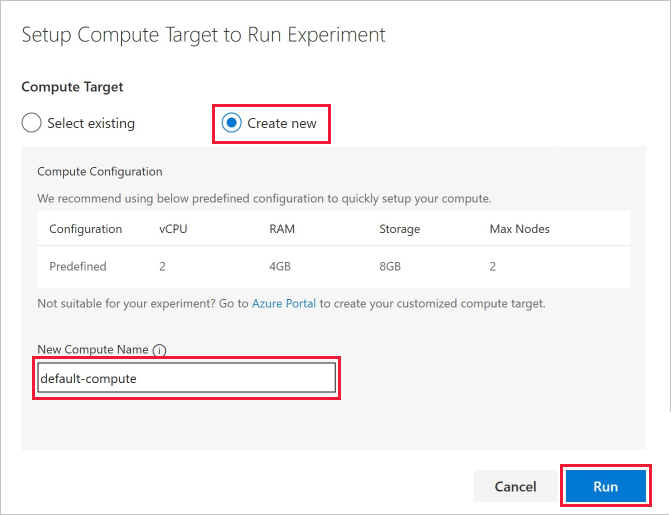


1. When the **Setup Compute Targets** dialog appears, if your workspace already has a compute resource, you can select it now. Otherwise, select **Create new**.

**Note**

The visual interface can only run experiments on Machine Learning Compute targets. Other compute targets will not be shown.

1. Provide a name for the compute resource.
2. Select **Run**.



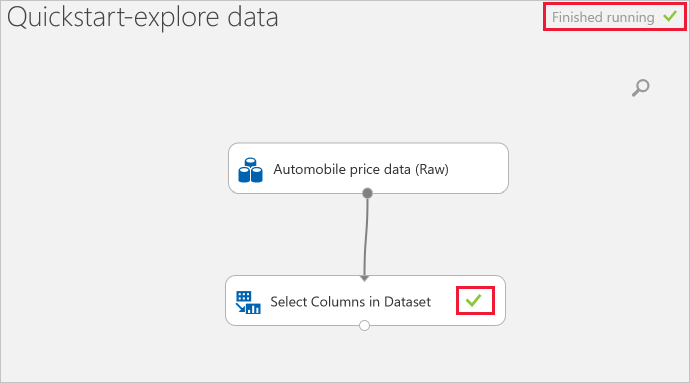
The compute resource will now be created. View the status in the top-right corner of the experiment.

**Note**

It takes approximately 5 minutes to create a compute resource. After the resource is created, you can reuse it and skip this wait time for future runs.

The compute resource will autoscale to 0 nodes when it is idle to save cost. When you use it again after a delay, you may again experience approximately 5 minutes of wait time while it scales back up.

After the compute target is available, the experiment runs. When the run is complete, a green checkmark appears on each module.



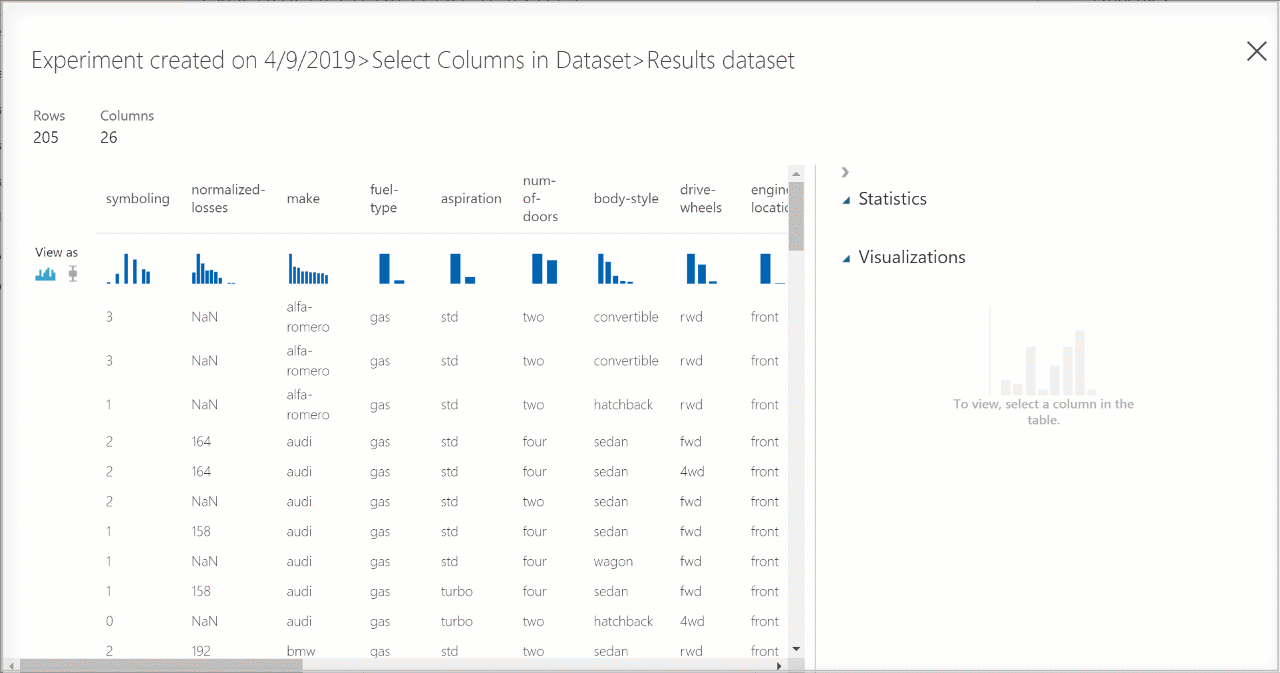
# Preview the data

Now that you have run your initial experiment, you can visualize the data to understand more about the information you have to work with.

1. Select the output port at the bottom of the **Select Columns in Dataset** then select **Visualize**.
2. Click on different columns in the data window to view information about that column.

In this dataset, each row represents an automobile, and the variables associated with each automobile appear as columns. There are 205 rows and 26 columns in this dataset.

Each time you click a column of data, the **Statistics** information and **Visualization** image of that column appears on the left. For example, when you click on **num-of-doors** you see it has 2 unique values and 2 missing values. Scroll down to see the values: two and four doors.



1. Click on each column to understand more about your dataset.

# Prepare data

A dataset usually requires some preprocessing before it can be analyzed. You might have noticed the missing values present in the columns of various rows. These missing values need to be cleaned so the model can analyze the data correctly. You'll remove any rows that have missing values. Also, the **normalized-losses** column has a large proportion of missing values, so you'll exclude that column from the model altogether.

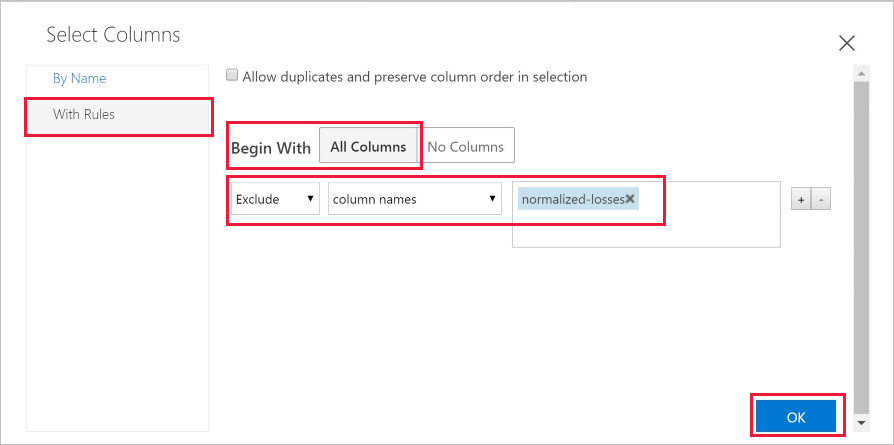
**Tip**

Cleaning the missing values from input data is a prerequisite for using most of the modules.

# Remove column

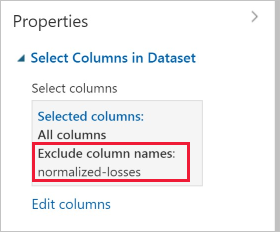
First, remove the **normalized-losses** column completely.

1. Select the **Select Columns in Dataset** module.
2. In the **Properties** pane to the right of the canvas, select **Edit columns**.
   * Leave **With rules** and **ALL COLUMNS** selected.
   * From the drop-downs, select **Exclude** and **column names**, and then click inside the text box. Type **normalized-losses**.
   * On the lower right, select **OK** to close the column selector.



Now the properties pane for Select Columns in Dataset indicates that it will pass through all columns from the dataset except **normalized-losses**.

The properties pane shows that the **normalized-losses** column is excluded.



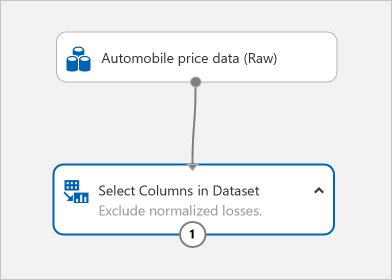
You can add a comment to a module by double-clicking the module and entering text. This can help you see at a glance what the module is doing in your experiment.

1. Double-click the **Select Columns in Dataset** module and type the comment "Exclude normalized losses."

After you type the comment, click outside the module. A down-arrow appears to show that the module contains a comment.

1. Click on the down-arrow to display the comment.

The module now shows an up-arrow to hide the comment.



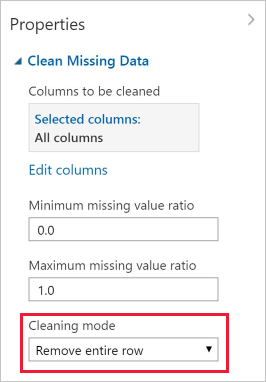
### Clean missing data

Now add another module that removes any remaining row that has missing data.

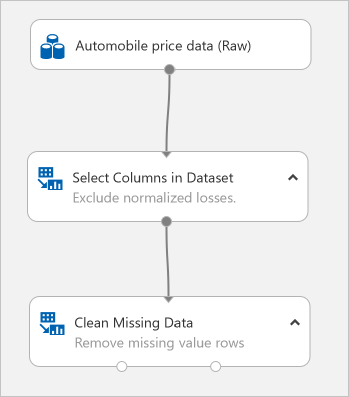
1. Type **Clean** in the Search box to find the **Clean Missing Data** module.
2. Drag the **Clean Missing Data** module to the experiment canvas and connect it to the **Select Columns in Dataset** module.
3. In the Properties pane, select **Remove entire row** under **Cleaning mode**.

These options direct **Clean Missing Data** to clean the data by removing rows that have any missing values.

1. Double-click the module and type the comment "Remove missing value rows."



Your experiment should now look something like this:



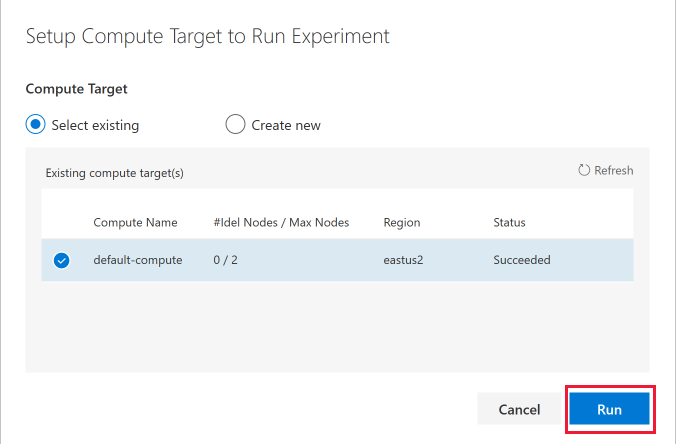
## Visualize the results

Since you made changes to the modules in your experiment, the status has changed to "In draft". To visualize the new clean data, you first have to run the experiment again.

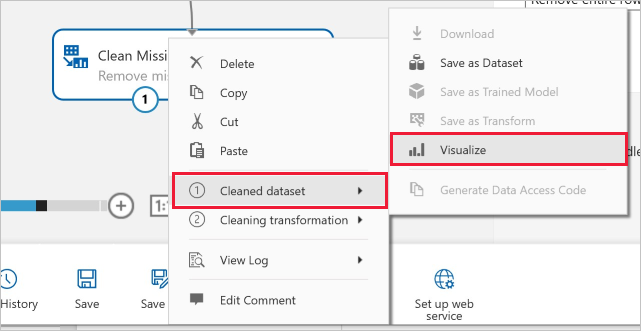
1. Select **Run** at the bottom to run the experiment.

This time you can reuse the compute target you created earlier.

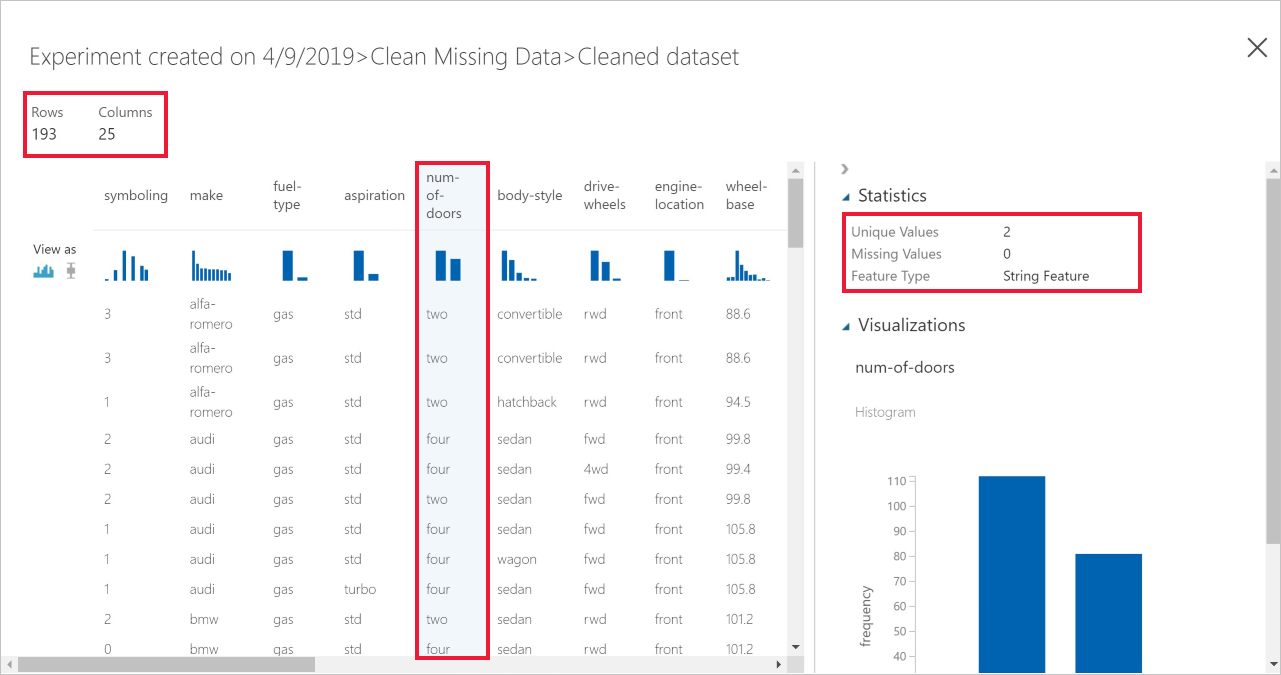
1. Select **Run** in the dialog.



1. When the run completes, right-click on the **Clean Missing Data** module to visualize the new clean data.



1. Click on different columns in the cleaned data window to see how data has changed.



There are now 193 rows and 25 columns.

When you click on **num-of-doors** you see it still has 2 unique values but now has 0 missing values.