Data Transformation

Citizen Analytics – An Initiative by Data Science Team

START ▸

Learning Objectives

By the end of this module, you will be able to:

01 Describe steps to perform data manipulation in Azure ML.

02 Describe how to use partition, sample and split data in Azure ML.

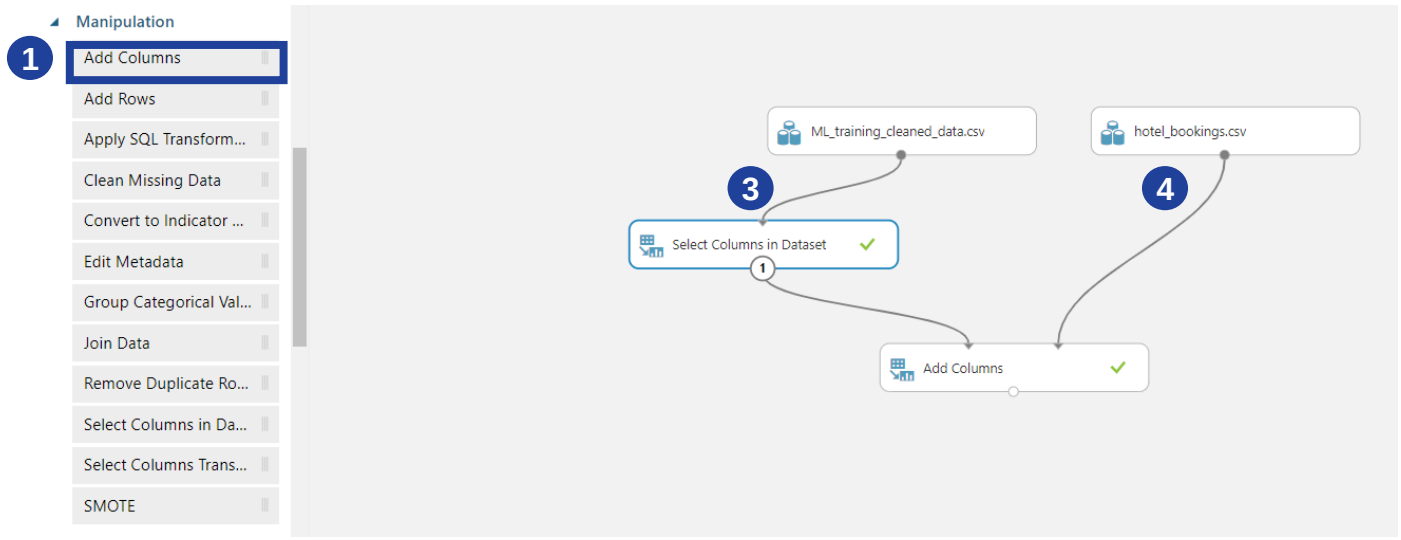
Content

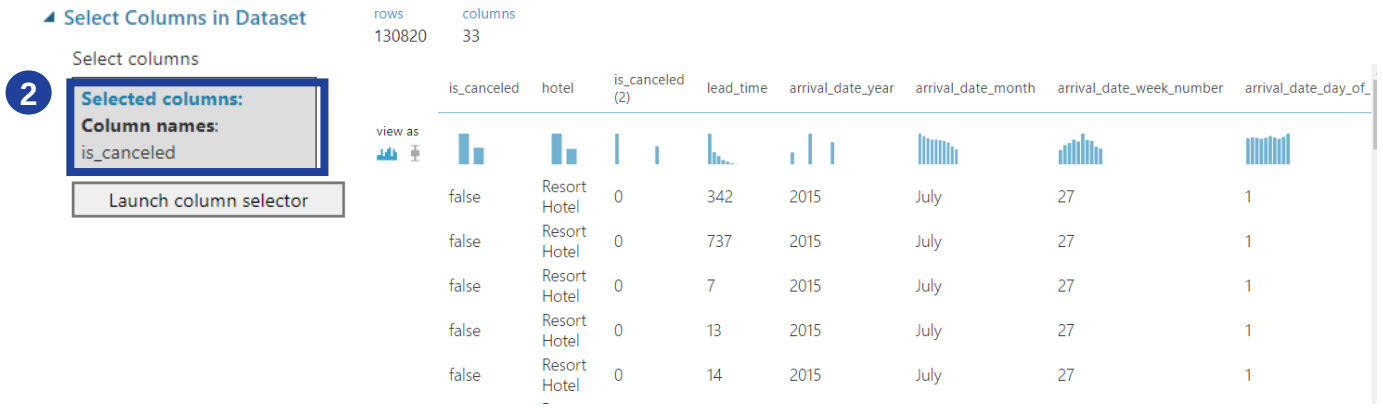
Data Manipulation

Data Manipulation - Add Columns

1. Add two datasets and select **“Add Columns”**
2. Add **“Select Columns in Dataset”** to select column “is_cancelled” from dataset 1 - “ML_training_cleaned_data.csv”
3. Connect dataset 1 to **“Select Columns in Dataset”** then to **“Add Columns”**
4. Connect dataset 2 - “hotel bookings.csv” to “Add Columns”
5. Click **“Run”** to obtain the combined dataset.

Use “Add Columns” to concatenate 2 datasets

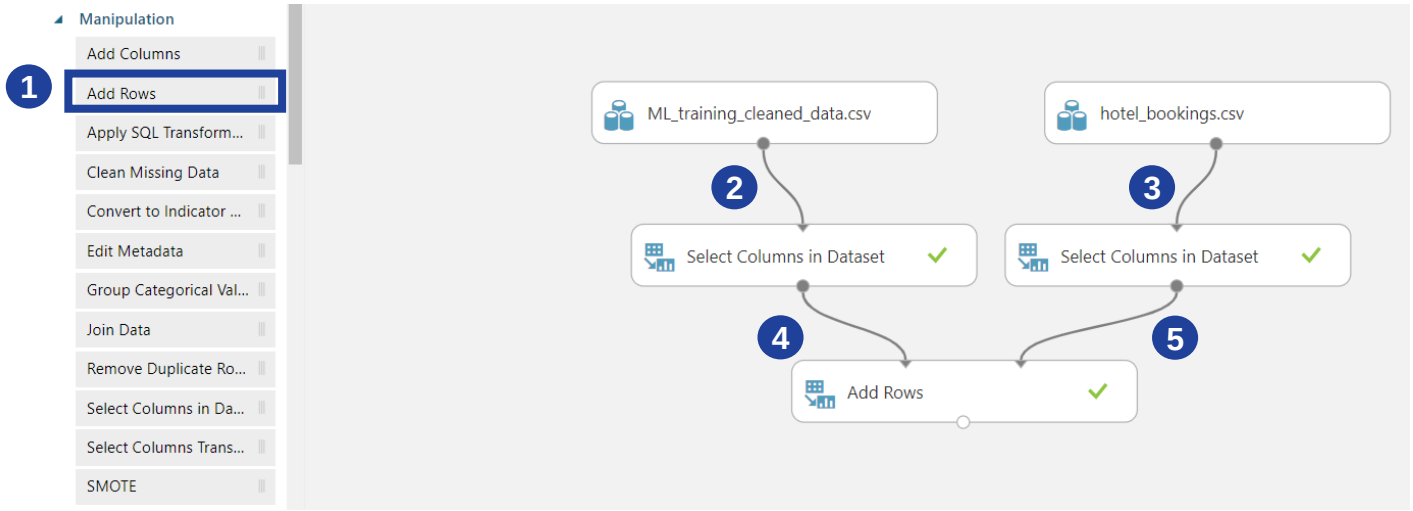




Data Manipulation - Add Rows

1. Add two datasets and select **“Add Rows”.**
2. Add **“Select Columns in Dataset”** to select column required to be appended from dataset 1 - “ML_training_cleaned_data.csv”.
3. Add **“Select Columns in Dataset”** to select column required to be appended from dataset 2 - “hotel bookings.csv”.
4. Connect dataset 1 to left port of **“Add Columns”.**
5. Connect dataset 2 to the right port of **“Add Columns”.** The dataset to append should be connected to the second (right) port.
6. Click **“Run”** to the obtain appended dataset. The number of rows in the output dataset should equal the sum of the rows of both input datasets.

Use “Add Rows” to concatenate 2 datasets



Experiment created on 6/29/2020 › ML_training_cleaned_data.csv › dataset

rows	columns
130820	34

Experiment created on 6/29/2020 › hotel_bookings.csv › dataset

rows	columns
119390	32

Experiment created on 6/29/2020 › Add Rows › Results dataset

rows	columns
250210	2

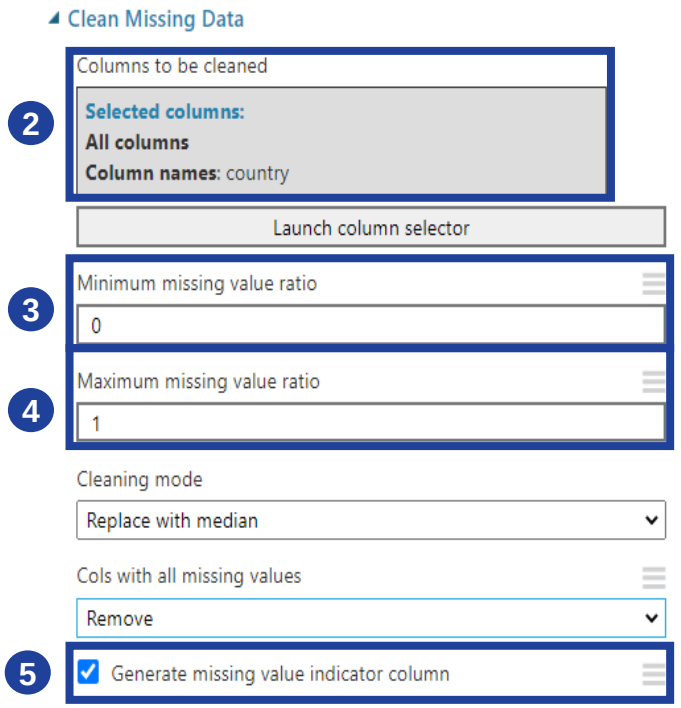
view as

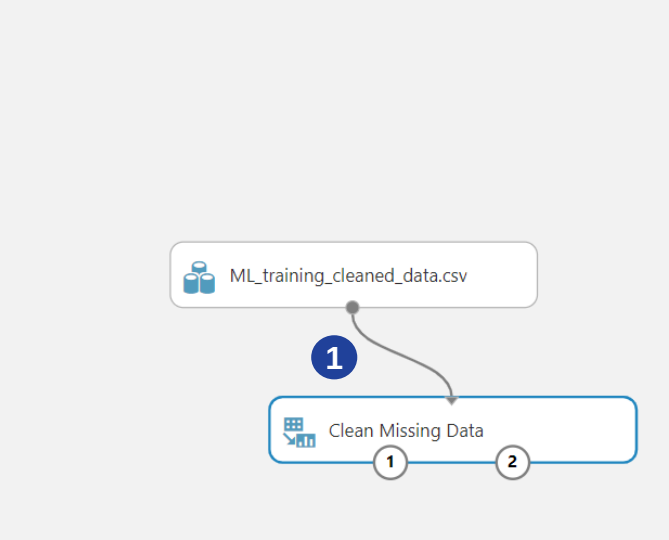
is_canceled	country
0	PRT
0	PRT
0	GBR
0	GBR

Data Manipulation - Clean Missing Data

1. Add **"Clean Missing Data"** and connect to a dataset.
2. Choose columns that contain missing values to be treated in **"Column to be cleaned".** Multiple columns selection is possible, but must use same replacement methods for all the selected columns.
3. Specify a minimum number of missing values to perform treatment in **"Minimum missing value ratio".** By default, this is set to 0, which means missing values are treated even if there is only one missing value.
4. Specify a maximum number of missing values to perform treatment in **"Maximum missing value ratio".** For example, 0.3 indicates missing values are treated if 30% or lesser rows contain missing values, otherwise if more than 30%, the values will not be treated. Can be used in combination with minimum missing value ratio to set upper and lower bound threshold to execute the missing treatment.
5. **"Generate missing value indicator column"**: Tick this option to indicate whether the values in the column met the criteria for missing value cleaning and examine they're treated accordingly.

Use "Clean Missing Data" to fill missing values



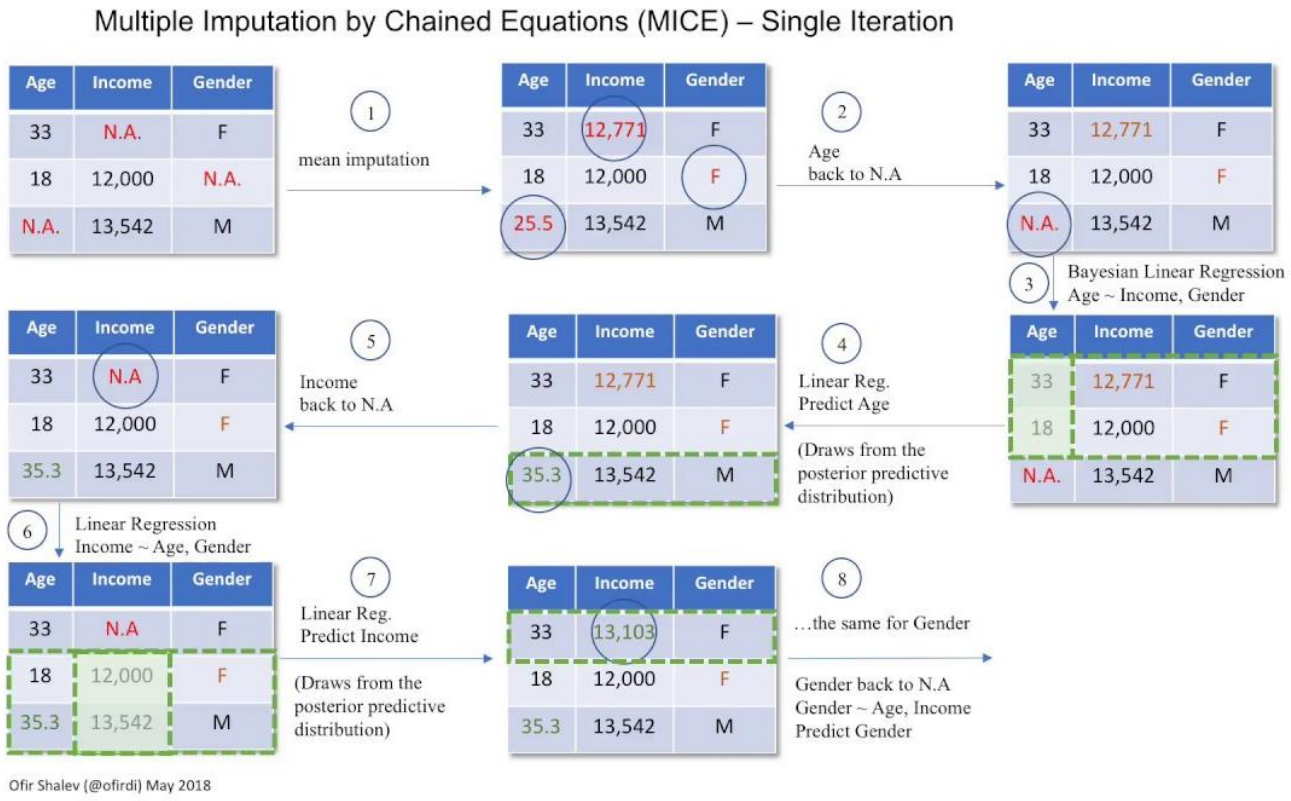


Data Manipulation - Clean Missing Data

“Cleaning Mode”

- **Replace using MICE:** Also known as "Multivariate Imputation using Chained Equations”, each variable with missing data is modeled conditionally using the other variables in the data to fill in the missing values.
- **Custom substitution value**: Fill all missing values with a placeholder such as 0 or NA.
- **Replace with mean**: Fill missing values with (of) each column with column mean.
- **Replace with median**: Fill missing values with (of) each column with column median.
- **Replace with mode**: Fill missing values with (of) each column with column mode.
- **Remove entire row**: Remove any row with one or more missing values, useful for a randomly missing value.
- **Remove entire column**: Remove any column with one or more missing values, useful for a randomly missing value.
- **Replace using Probabilistic PCA**: Replaces the missing values by using a linear model that analyzes the correlations between the columns and estimates a low-dimensional approximation of the data, from which the full data is reconstructed.

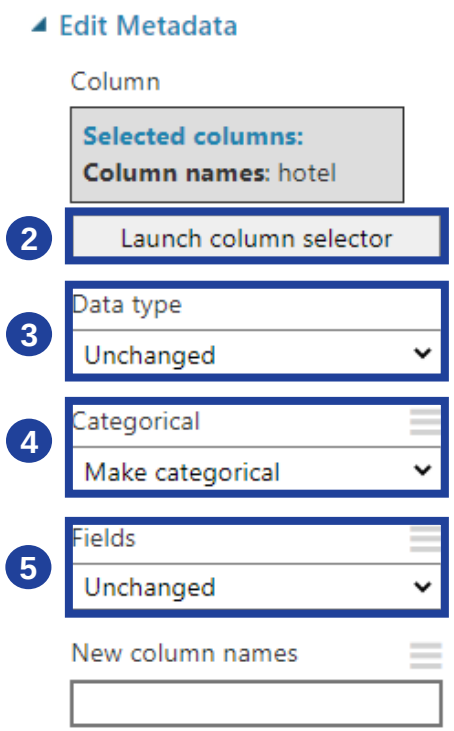
Graphical explanation on “MICE”

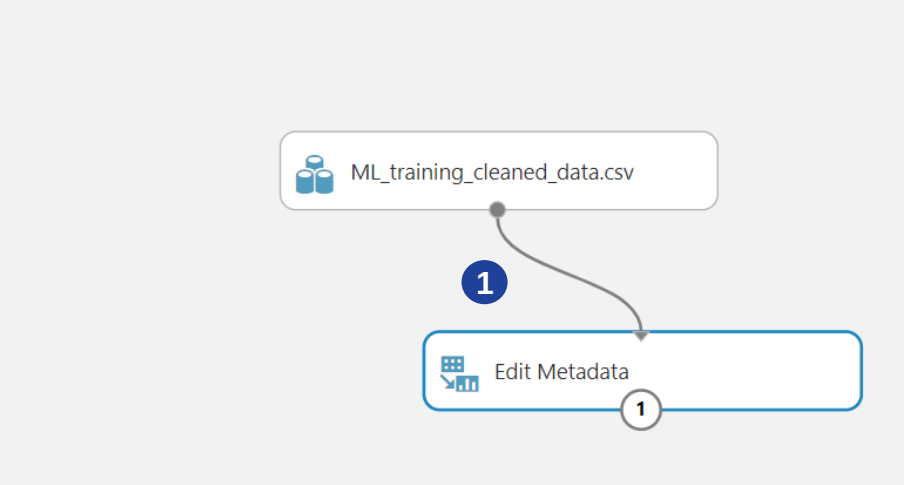


Data Manipulation - Edit Metadata

1. Add **"Edit Metadata"** and connect to a dataset.
2. Launch **"Launch column selector"** to choose a column or multiple columns to work with.
3. **Data type** : assign a different data type to the selected columns.
   - String, Integer, Floating point, Boolean, DateTime, and TimeSpan.
   - Metadata changes apply to all selected columns ONLY.
   - Changes (of) in data type affect how the data is handled in downstream operations.
4. **Categorical** : specify values in the selected columns to be treated as categories.
5. **Fields** : specify how Azure ML treats column/s in modelling
   - **Feature:** flag column/s as feature
   - **Label:** flag column/s as target label
   - **Weight:** flag a numeric column as weights in machine learning scoring or training operations
   - **Clear feature:** remove the feature flag
   - **Clear label:** remove the label flag
   - **Clear weight:** remove the weight flag

Use "Edit Metadata" to manipulate data type

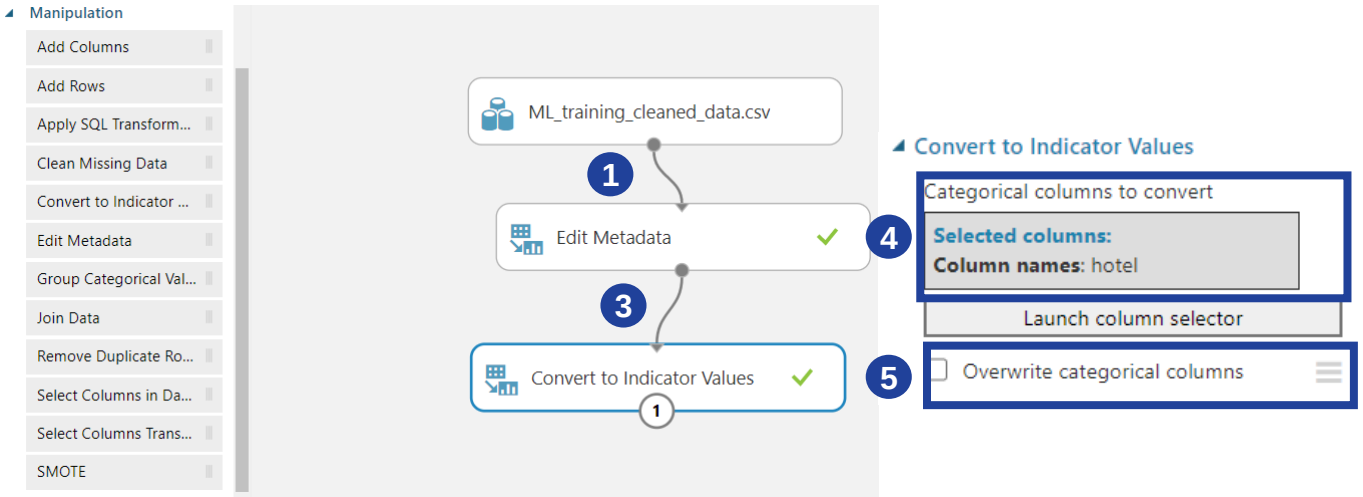




Data Manipulation - Convert to Indicator Values

1. Add **"Edit Metadata"** and connect to a dataset.
2. Launch **"Launch column selector"** to choose a column or multiple columns to make it categorical, for example "hotel".
3. Add **"Convert to Indicator Values"** and connect to "**Edit Metadata". "Convert to Indicator Values"** is used for one-hot-encoding, convert columns with categorical values into a series of binary indicator columns that can more easily be used as features in a machine learning model.
4. At "**Categorical columns to convert",** choose one or more categorical columns, for example, "hotel".
5. Select the **"Overwrite categorical columns"** if only the new Boolean columns are to be retained.
6. Click **"Run"** to obtain the new indicator columns.

Use "Convert to Indicator Values" to do one-hot-encoding



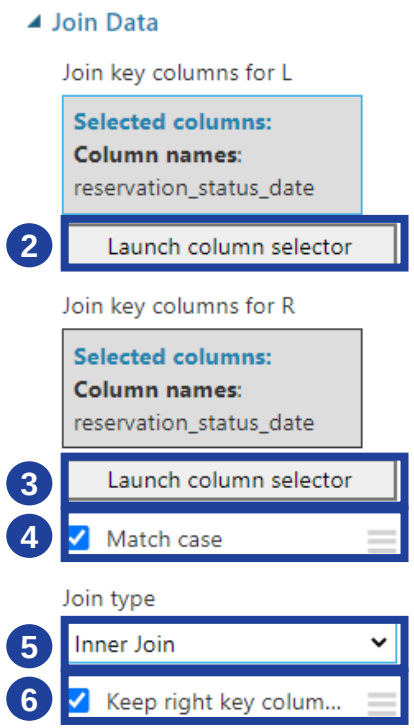
hotel
Resort Hotel
Resort Hotel
Resort Hotel
Resort Hotel

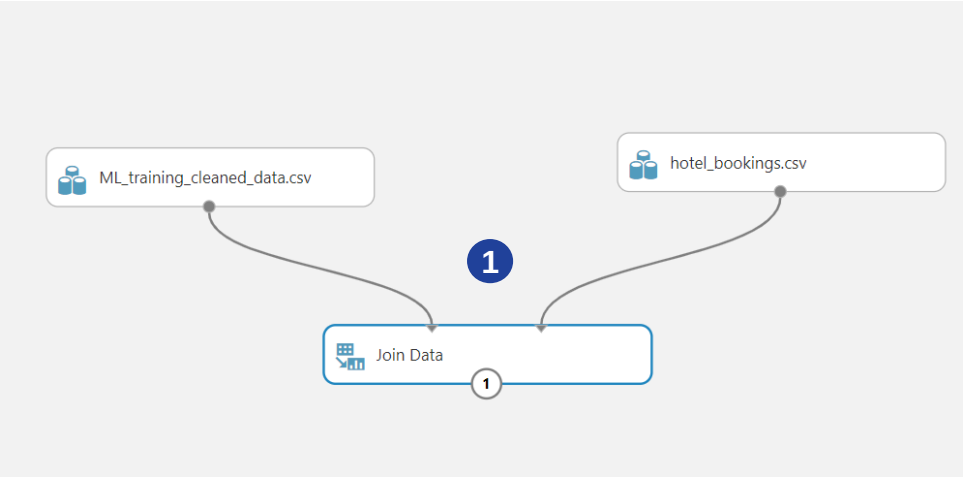
hotel-City Hotel	hotel-Resort Hotel
0	1
0	1
0	1
0	1

Data Manipulation - Join Data

1. Add two datasets and connect to **“Join Data”.**
2. Launch **“Launch column selector”** to choose a single key column for the dataset on the left input.
3. Launch **“Launch column selector”** to choose a single key column for the dataset on the right input.
4. Tick **“Match case”** to ensure case sensitivity when joining on a text column, for example “A123” will be considered different from “a123”.
5. **“Join type”** - how datasets are combined.
   - **“Inner Join”** - return combined rows only when a value in key columns are matched.
   - **“Left Outer Join”** - return combined rows for all rows from the left table. If a row in the left table has not matched any rows in the right table, the return row to the left table from the right will contain missing values.
   - **“Full Outer Join”** - return all rows from the left and right table.
   - **“Left Semi-Join”** - return only rows from the left table when key columns are matched.
6. Tick **“Keep right key columns in joined table”** to return keys from both input tables.
7. Click **“Run”** to obtain the combined dataset.

Use “Join Data” to merge data

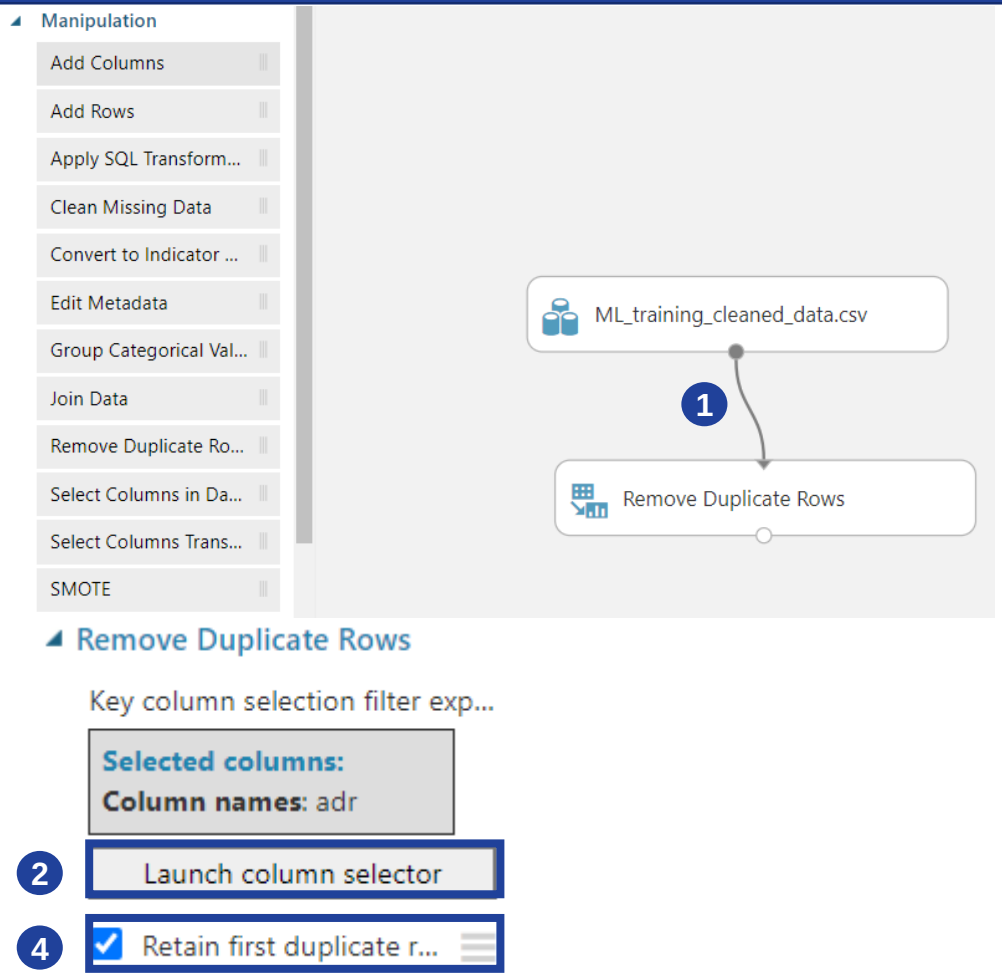




Data Manipulation - Remove Duplicate Rows

1. Add dataset and connect to **“Remove Duplicate Rows”.**
2. Under **“Key column selection filter expression”,** launch **“Launch column selector”** to choose column(s) to identify duplicate records.
3. If one column is selected, it will be used to find duplicate rows. Whereas if multiple columns are selected, the value combinations of the selected columns will be used to find duplicate rows.
4. Tick **“Retain first duplicate row”** to indicate which row to return when duplicates are found:
   - If selected, the first row is returned, and others discarded.
   - If not, the last duplicate row is kept in the results, and others are discarded.
5. Click **“Run”** to obtain no duplicate datasets

Use “Remove Duplicate Rows” to remove duplicate rows



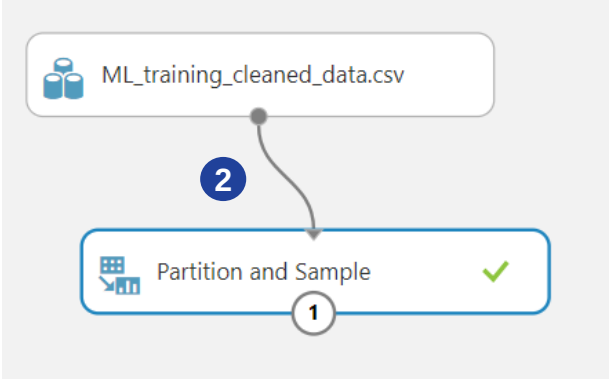
Sample and Split

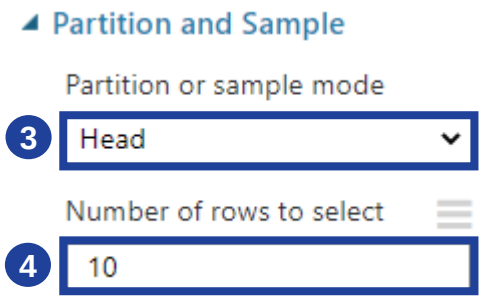
Sample and Split - Partition and Sample

"**Partition and Sample**" can be used for below purposes:

- Divide data into multiple smaller datasets of the same size for cross-validation.
- Separate data into groups and then working with data from a specific group. After randomly assigning cases to different groups, modify the features that are associated with only one group.
- Sampling. Extract a percentage of the data, apply random sampling, or choose a column to use for balancing the dataset and perform stratified sampling on its values.
- Create a smaller dataset for testing.

Use "Partition and Sample" to slice or resample data





rows 10 columns 34

hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in
Resort Hotel	false	320	2015	July	27	1	0
Resort Hotel	false	320	2015	July	27	1	0
Resort Hotel	false	7	2015	July	27	1	0
Resort Hotel	false	13	2015	July	27	1	0
Resort Hotel	false	14	2015	July	27	1	0
Resort Hotel	false	14	2015	July	27	1	0
Resort Hotel	false	0	2015	July	27	1	0
Resort Hotel	false	9	2015	July	27	1	0
Resort Hotel	true	85	2015	July	27	1	0
Resort Hotel	true	75	2015	July	27	1	0

Get TOP N rows from a dataset

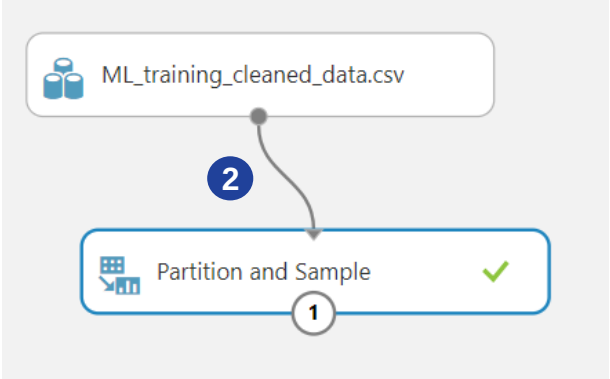
1. Return the first n rows, useful when to test an experiment on a small number of rows, and don't need the data to be balanced or sampled in any way.
2. Add dataset and connect to **"Partition and Sample".**
3. In "**Partition or sample mode",** Select **"Head".**
4. **"Number of rows to select":** type non-negative number for rows to return. If the number is larger than the total number of rows in the dataset, all rows are returned
5. Click **"Run"** to obtain the result.

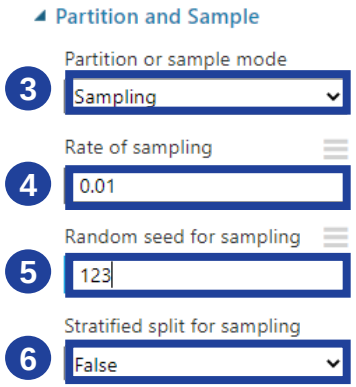
Sample and Split - Partition and Sample

Create a sample of data

1. Do simple random sampling or stratified random sampling. Use to create a smaller representative sample dataset for testing.
2. Add dataset and connect to **“Partition and Sample”.**
3. In “**Partition or sample mode”,** Select **“Sampling”.**
4. **“Rate of sampling”**: Insert value between 0 and 1 to specify the percentage of rows from the source dataset to be returned as output. The rows are shuffled and selected. For example, 0.5 indicates that the sampling rate is 50%, which means return half of the datasets after shuffling
5. **“Random seed for sampling”:** Set seed for results repeatability.
6. **“Stratified split for sampling”:** Set to True if it is important that the rows in the dataset should be divided evenly by some key column before sampling. Then, select a single strata column to use when dividing the dataset. The rows in the dataset are then divided as follows:
   - All input rows are grouped (stratified) by the values in the specified strata column.
   - Rows are shuffled within each group.
   - Each group is selectively added to the output dataset to meet the specified ratio.
7. Click **“Run”** to obtain the result.

Use “Partition and Sample” to slice or resample data





rows 130820 columns 34

hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays
Resort Hotel	false	320	2015	July	27	1	0
Resort Hotel	false	320	2015	July	27	1	0
Resort Hotel	false	7	2015	July	27	1	0

rows 1308 columns 34

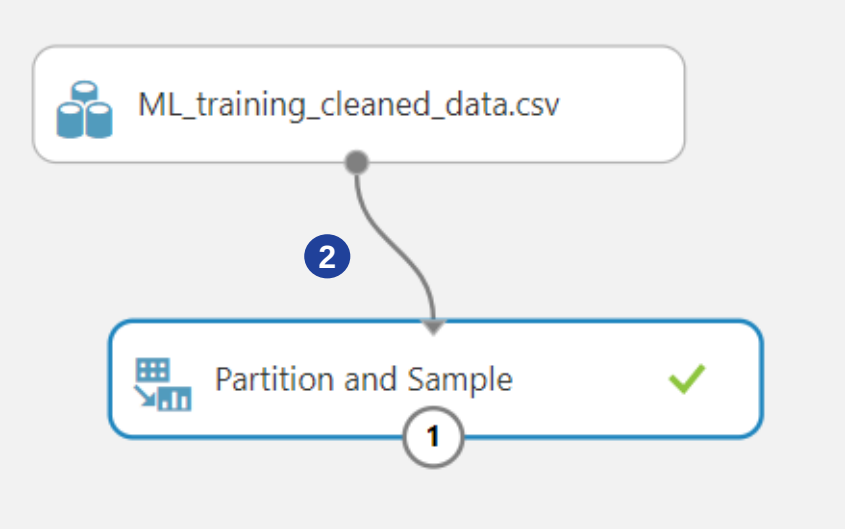
hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays
City Hotel	false	4	2016	October	43	19	0
City Hotel	false	211	2017	August	34	23	0
City Hotel	false	320	2016	September	38	13	0
City Hotel	false	17	2017	February	8	21	0

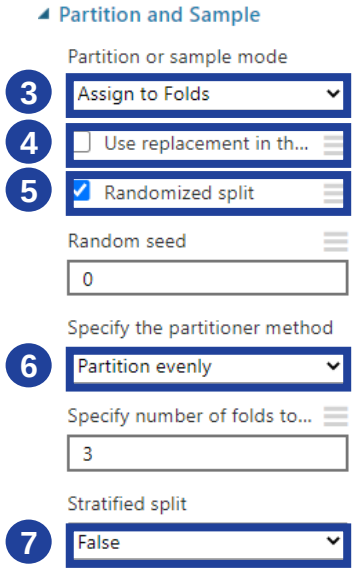
Sample and Split - Partition and Sample

Split data into partitions

1. Divide the dataset into subsets, useful when to create number of folds for cross-validation, or to split rows into several groups.
2. Add dataset and connect to **"Partition and Sample".**
3. In "**Partition or sample mode",** Select **"Assign to Folds".**
4. **"Use replacement in the partitioning" :** Tick this option for a sampled row to be reused into multiple partitioned folds.
5. **"Randomized split" :** Tick this option for rows to be randomly assigned to folds.
6. **"Specify the partitioner method": Indicate how data will be apportioned to each partition, using the below options:**
   - **Partition evenly:** To assign an equal number of rows in each partition. Type a whole number in the **"Specify number of folds to split evenly into text box".**
   - **Partition with customized proportions:** To specify the size of each partition as a comma-separated list. For example, to create three partitions, with the first partition containing 50% of the data, and the remaining two partitions each containing 25% of the data, click the List of proportions separated by the comma text box, and type these numbers: .5, .25, .25.
   - The sum of all partition sizes must add up to exactly 1.
7. **"Stratified split for sampling":** Set to True data to be stratified by key strata column before the split.
8. Click **"Run"** experiment.

Use "Partition and Sample" to slice or resample data



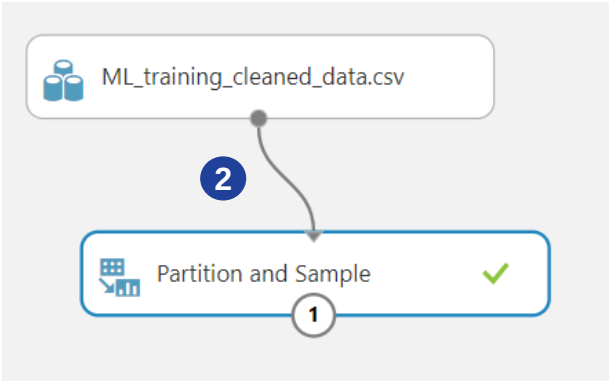


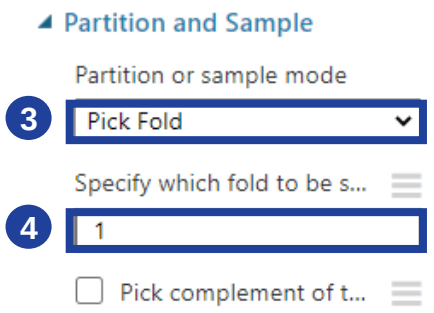
Sample and Split - Partition and Sample

Use data from a predefined partition

1. To load each partition for further analysis or processing after dividing a dataset into multiple partitions.
2. Connect **"Partition and Sample"** to output from previous **"Partition and Sample"**
3. In "**Partition or sample mode",** Select **"Pick Fold".** The previous **"Partition and Sample"** must (has) have **"Assign to Folds"** selected to generate some folds.
4. **"Specify which fold to be sampled from":** Select a partition to use by typing its index. Partition indices are 1-based. For example, if divided the dataset into three parts in the previous slide, the partitions would have the indices 1, 2, and 3.
5. One **"Partition and Sample"** module for one individual fold. Need to use (multiple) same multiple modules for taking in multiple folds for further analysis or processing
6. Click **"Run"** to obtain the result.

Use "Partition and Sample" to slice or resample data





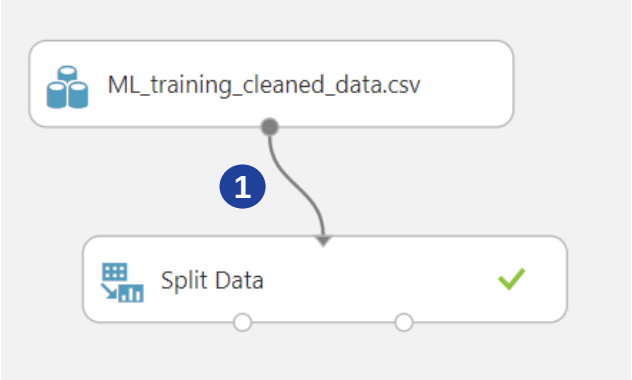
rows 43606 columns 34

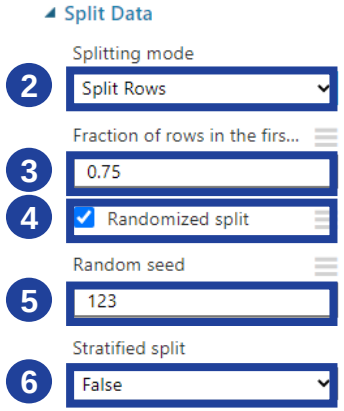
hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month
City Hotel	true	258	2015	July	27	2
City Hotel	false	262	2017	January	2	9
City Hotel	true	0	2016	March	13	23

Sample and Split - Split Data

1. Add dataset and connect to **“Split Data”.**
2. **“Splitting mode”**: **Split Rows**: Use this option to divide the data into two parts. By default, the data is divided 50-50.
3. **“Fraction of rows in the first output dataset”.** Use this option to determine how many rows go into the first (left-hand) output. All other rows will go to the second (right-hand) output. A value between 0 to 1 to be inserted. For example, 0.75 means the dataset will be split by 75:25 ratio.
4. Tick **“Randomized split”** to randomly select data into the two groups.
5. **“Random seed”:** Set seed for results repeatability.
6. **“Stratified split for sampling”:** Set to True to ensure two output datasets contain a representative sample of the values in the strata column selected.
7. Click **“Run”** to obtain the result.

Use “Split Data” to separate data into two subsets





rows 98115 columns 34

hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays
City Hotel	true	3	2015	September	36	5	2
City Hotel	true	259	2016	May	19	5	0
Resort Hotel	true	44	2016	August	35	25	2

rows 32705 columns 34

hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays
City Hotel	false	2	2017	March	10	11	1
Resort Hotel	false	10	2016	December	49	3	0
City Hotel	false	175	2016	August	32	3	3

Summary

Summary

1 Data manipulation in Azure ML

- You learnt to add columns and rows, clean missing data, assign data type, perform one-hot encoding, and remove duplicated values.
- https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/data-transformation-manipulation

2 Data partition and split in Azure ML

- You learnt 3 modes of data partition in Azure ML: Sampling, Assign to Folds, and Pick Fold
- You learnt to split data by rows in Azure ML
- https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/data-transformation-sample-and-split

Thank you for your passion!

PETRONAS