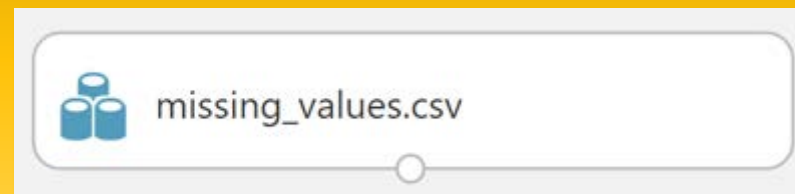


Missing value handling

MISSING VALUE HANDLING



Missing value handling

In this session

1. Replace missing values with the mean
2. Replace missing values with the median
3. Replace missing values with an interpolated estimate
4. Replace missing values with a constant
5. Replace missing values using imputation
6. Replace missing values with a missing rank
7. Replace missing values with a dummy
8. Replace missing values with 0
9. Create an indicator variable for "missing."
10. Replace missing values with a string
11. Add an indicator variable showing which strings are considered "missing."
12. Delete columns that are missing too many values to be useful
13. Delete rows that are missing critical values

Missing value handling
























We need data that is:

- Relevant
- Connected
- Accurate
- Enough to work with



Missing value handling

Example of missing values dataset

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
								
Tony	48	27		1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick		17		4				
Bruce	37	14	63		1	veggie		NA
Steve	83		77	7	1	chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	162	5	3			-
Carol		3	127	11	1	veggie	1	""
Mandy	44	2	68	8	1	chicken		null

Missing value handling

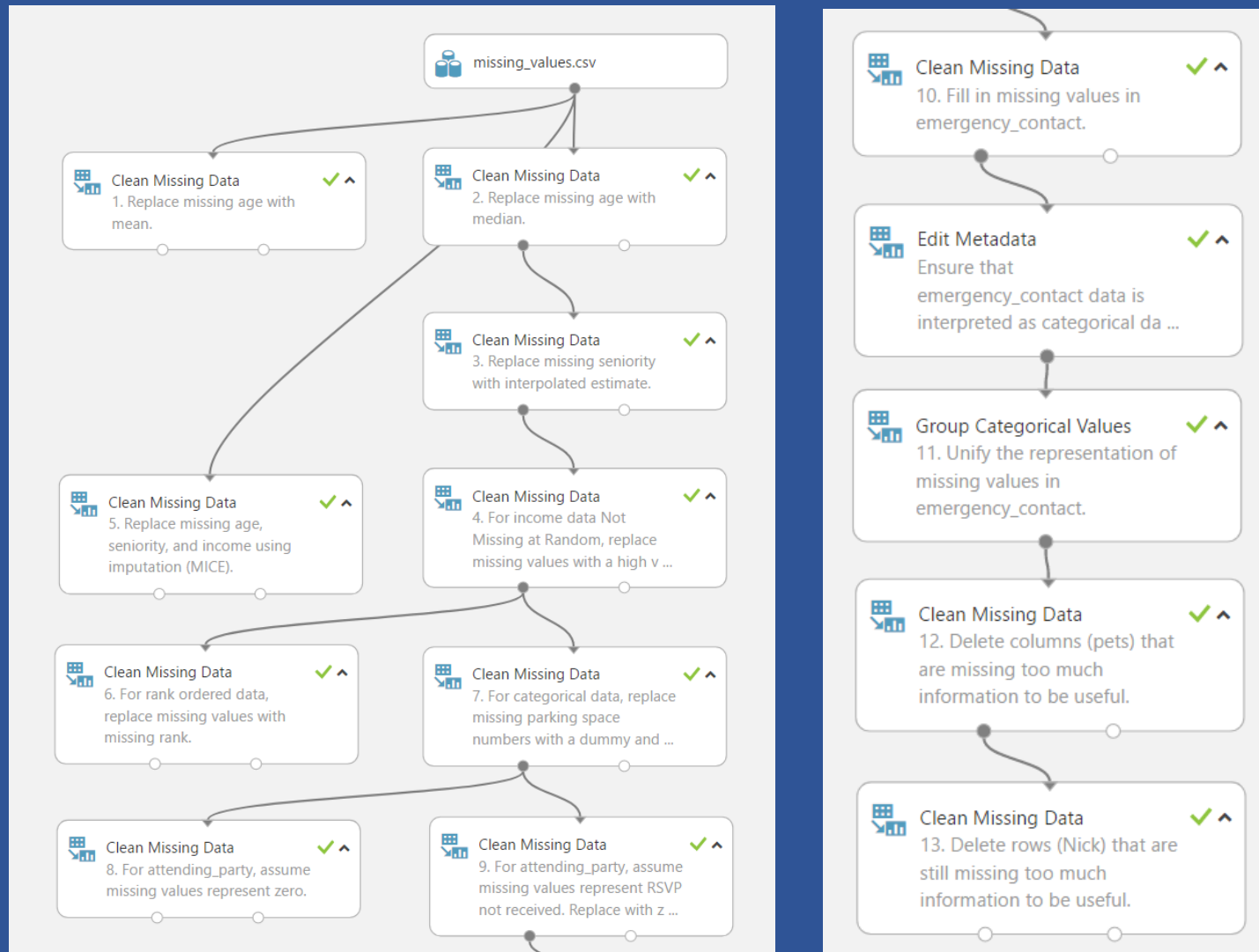
Example of missing values dataset CSV file

missing_values.csv

	A	B	C	D	E	F	G	H	I
1		age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
2	Tony	48	27		1	5	shrimp		Pepper
3	Donald	67	25	86	10	2	beef		Jane
4	Henry	69	21	95	6	1	chicken	62	Janet
5	Janet	62	21	110	3	1	beef		Henry
6	Nick		17		4				
7	Bruce	37	14	63		1	veggie		NA
8	Steve	83		77	7	1	chicken		n/a
9	Clint	27	9	118	9		shrimp	3	None
10	Wanda	19	7	52	2	2	shrimp		empty
11	Natasha	26	4	162	5	3			_
12	Carol		3	127	11	1	veggie	1	""
13	Mandy	44	2	68	8	1	chicken		null

Missing value handling

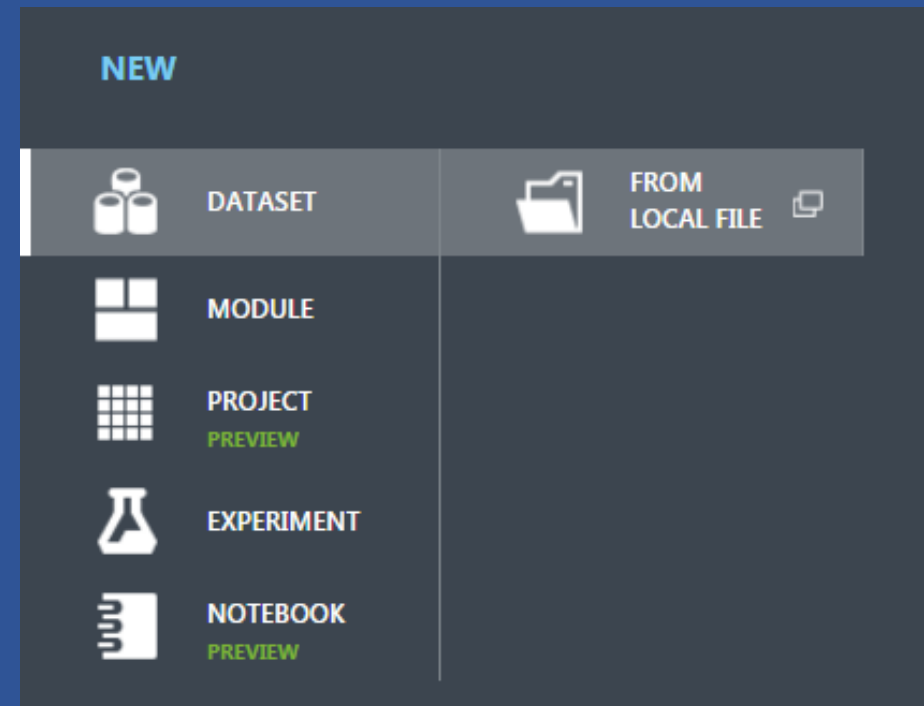
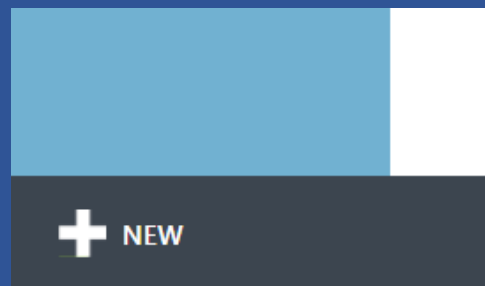
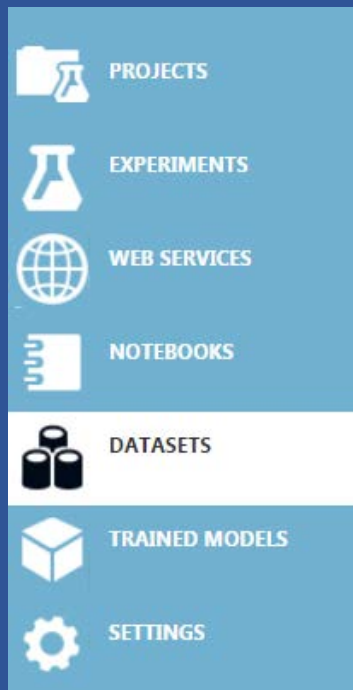
Experiment: Methods for handling missing values



Missing value handling

Import local data file to Azure ML Studio Dataset

1. Select Datasets tab from menu
2. Click New (+) at the button left corner
3. Click “From local file”



Missing value handling

4. Choose file and type description

×

Upload a new dataset

SELECT THE DATA TO UPLOAD:

Choose File

missing_values.csv

☐ This is the new version of an existing dataset

ENTER A NAME FOR THE NEW DATASET:

missing_values.csv

SELECT A TYPE FOR THE NEW DATASET:

Generic CSV File with a header (.csv) ▼

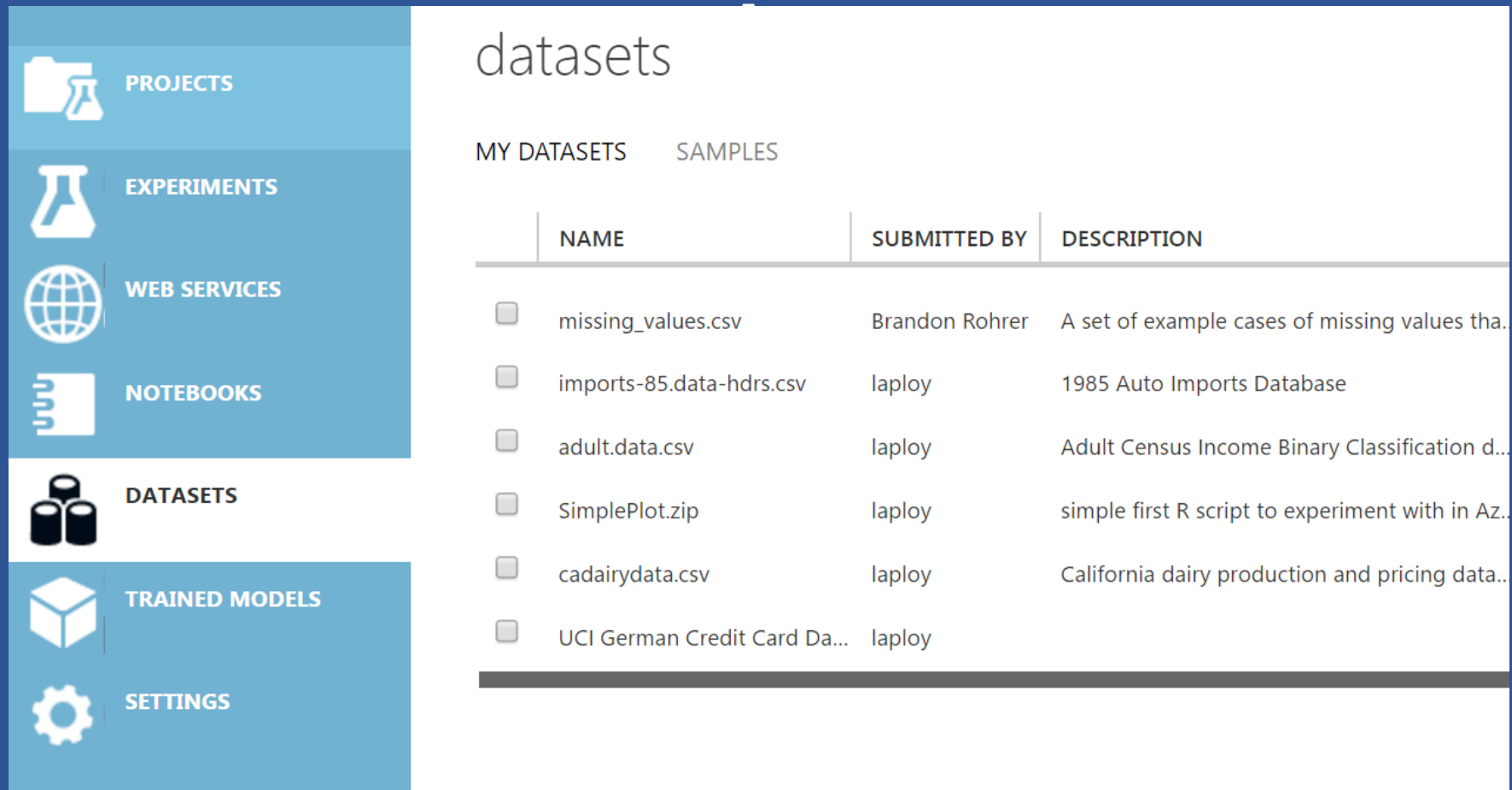
PROVIDE AN OPTIONAL DESCRIPTION:

sample of missing data

✓

Missing value handling

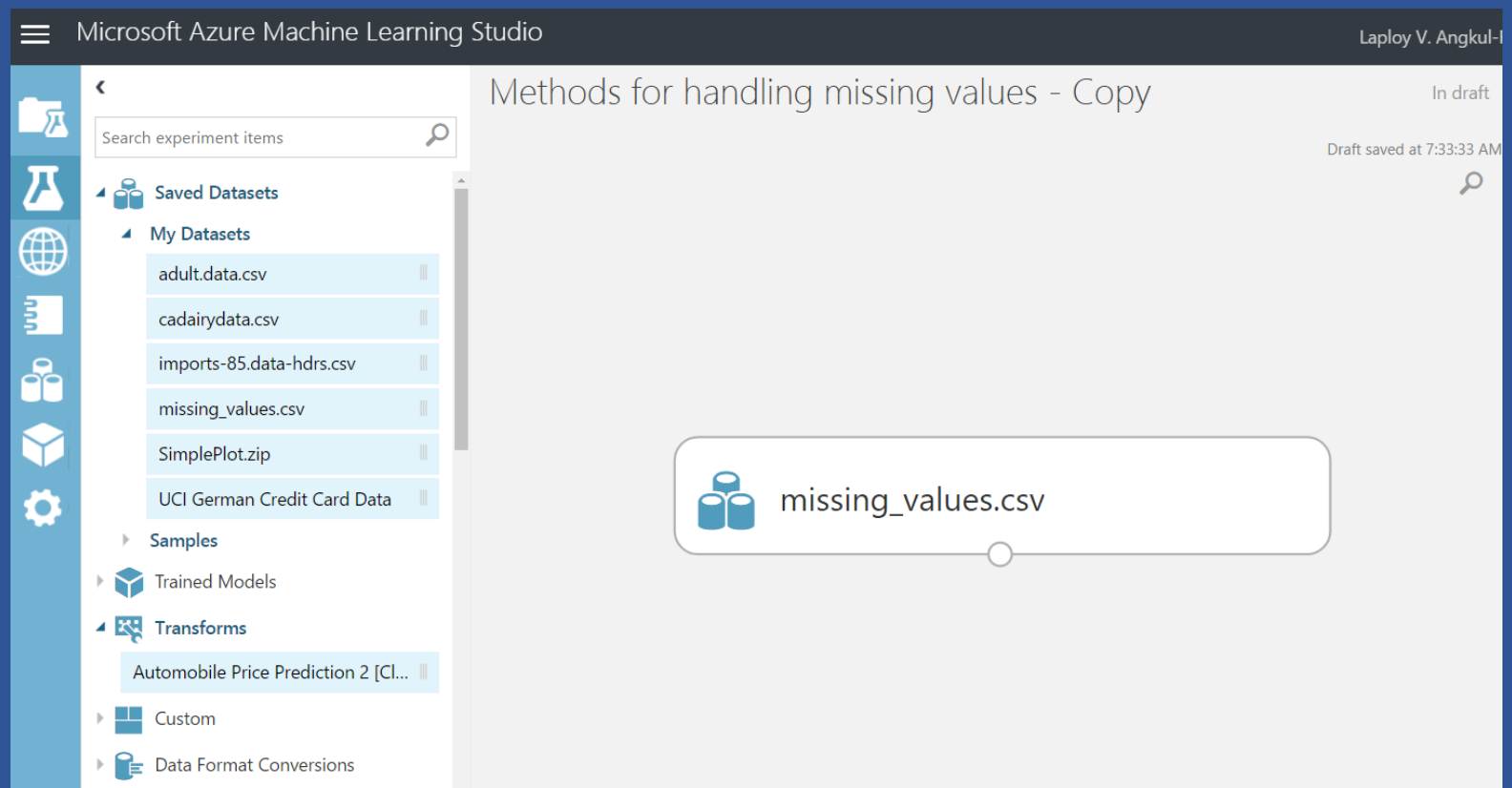
Saved datasets list



	NAME	SUBMITTED BY	DESCRIPTION
<input type="checkbox"/>	missing_values.csv	Brandon Rohrer	A set of example cases of missing values tha...
<input type="checkbox"/>	imports-85.data-hdrs.csv	laploy	1985 Auto Imports Database
<input type="checkbox"/>	adult.data.csv	laploy	Adult Census Income Binary Classification d...
<input type="checkbox"/>	SimplePlot.zip	laploy	simple first R script to experiment with in Az...
<input type="checkbox"/>	cadairydata.csv	laploy	California dairy production and pricing data...
<input type="checkbox"/>	UCI German Credit Card Da...	laploy	

Missing value handling

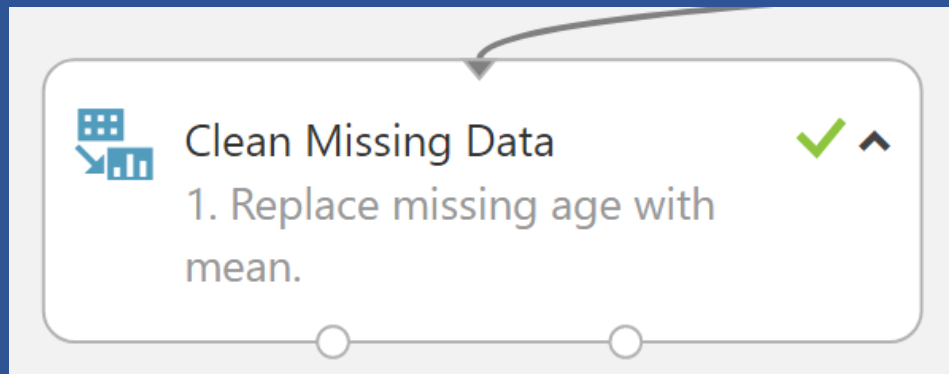
- Create new blank Experiment
- Select **missing_values.csv** from **Saved Datasets**
- Drag & drop into canvas



Missing value handling

Replace missing values with the mean

- Change project name to “Methods for handling missing values”
- Drag & drop Clean Missing Data module
- Select column **age**
- Configure “Cleaning mode” to **Replace with mean**
- Comment = 1. Replace missing age with mean.
- Run/Visualize



Properties Project

Clean Missing Data

Columns to be cleaned

Selected columns:
Column names: age

Launch column selector

Minimum missing value ...

0

Maximum missing value...

1

Cleaning mode

Replace with mean

Cols with all missing val...

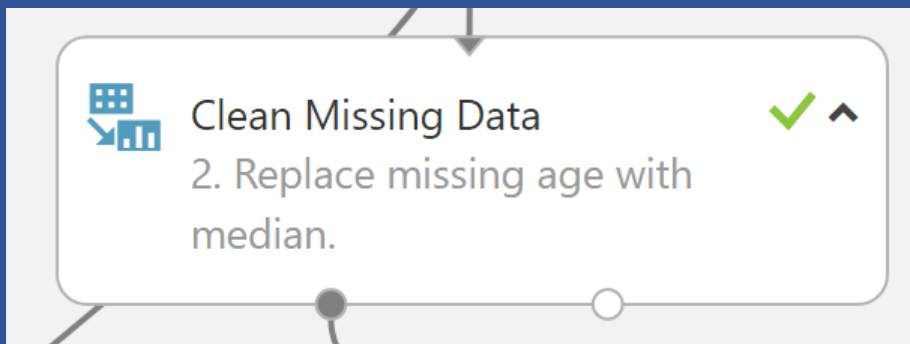
Remove

☐ Generate missing va...

Missing value handling

Replace missing values with the median

- Drag & drop Clean Missing Data module
- Select column **age**
- Configure “Cleaning mode” to **Replace with median**
- Comment = 2. Replace missing age with median.
- Run/Visualize



Properties Project

Clean Missing Data

Columns to be cleaned

Selected columns:
Column names: age

Launch column selector

Minimum missing value ...

Maximum missing value...

Cleaning mode
Replace with median ▼

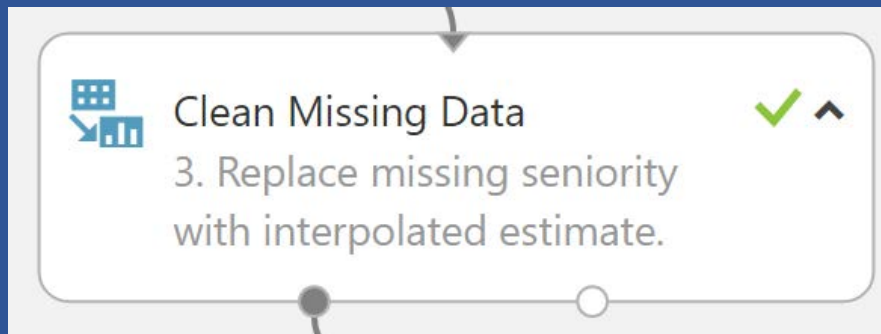
Cols with all missing val...

☐ Generate missing va...

Missing value handling

Replace missing values with an interpolated estimate

- Drag & drop Clean Missing Data module
- Select column **year_seniority**
- Configure “Cleaning mode” to **Custom substitution value**
- Set **Replacement value** to **11.5**
- Comment = 3. Replace missing seniority with interpolated estimate.
- Run/Visualize



Properties Project

Clean Missing Data

Columns to be cleaned

Selected columns:
Column names: years_seniority

Launch column selector

Minimum missing value ratio

Maximum missing value ratio

Cleaning mode
Custom substitution value

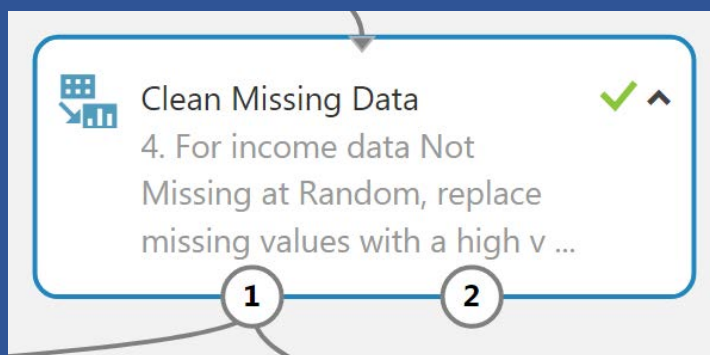
Replacement value

☐ Generate missing value indica..

Missing value handling

Replace missing values with a constant

- Drag & drop Clean Missing Data module
- Select column **income**
- Configure “Cleaning mode” to **Custom substitution value**
- Set **Replacement value** to **250**
- Comment = 4. For income data Not Missing at Random, replace missing values with a high value.
- Run/Visualize



Properties Project

Clean Missing Data

Columns to be cleaned

Selected columns:
Column names: income

Launch column selector

Minimum missing value ratio

Maximum missing value ratio

Cleaning mode
Custom substitution value

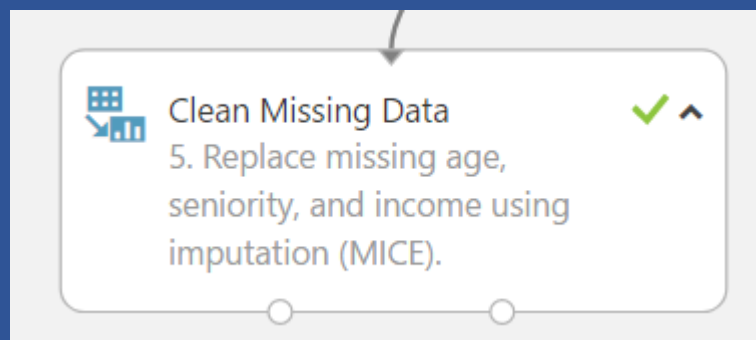
Replacement value

☐ Generate missing value indica..

Missing value handling

Replace missing values using imputation

- Drag & drop Clean Missing Data module
- Select column **years_seniority.age.income**
- Configure “Cleaning mode” to **Replace using MICE**
- Cols with all missing values = Remove
- Number of iterations = 5
- Comment = 5. Replace missing age, seniority, and income using imputation (MICE).
- Run/Visualize



Properties Project

Clean Missing Data

Columns to be cleaned

Selected columns:
Column names:
years_seniority,age,income

Launch column selector

Minimum missing value ratio

Maximum missing value ratio

Cleaning mode
Replace using MICE

Cols with all missing values
Remove

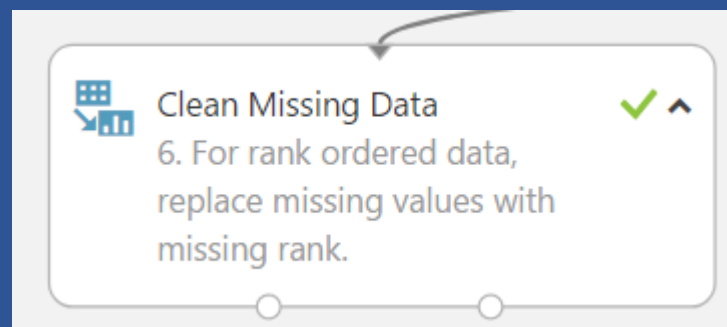
☐ Generate missing value indic...

Number of iterations

Missing value handling

Replace missing values with a missing rank

- Drag & drop Clean Missing Data module
- Select column **parking_space**
- Configure “Cleaning mode” to **Custom substitution value**
- Replacement value = 12
- Comment = 6. For rank ordered data, replace missing values with missing rank.
- Run/Visualize



Properties Project

Clean Missing Data

Columns to be cleaned

Selected columns:
Column names: parking_space

Launch column selector

Minimum missing value ratio

Maximum missing value ratio

Cleaning mode
Custom substitution value

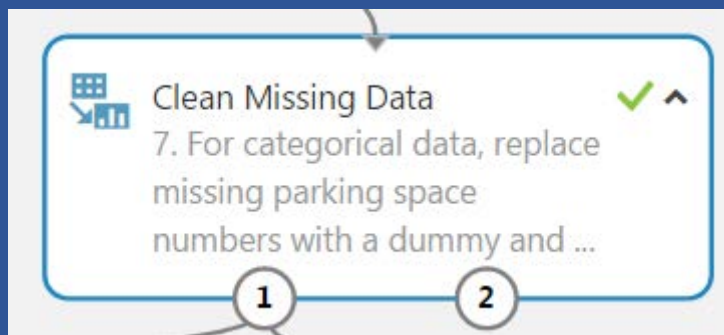
Replacement value

☐ Generate missing value indicato...

Missing value handling

Replace missing values with a dummy

- Drag & drop Clean Missing Data module
- Select column **parking_space**
- Configure “Cleaning mode” to **Custom substitution value**
- Replacement value = -99
- Comment = 7. For categorical data, replace missing parking space numbers with a dummy and include a missing values indicator column
- Run/Visualize



Properties Project

Clean Missing Data

Columns to be cleaned

Selected columns:
Column names: parking_space

Launch column selector

Minimum missing value ratio

Maximum missing value ratio

Cleaning mode
Custom substitution value

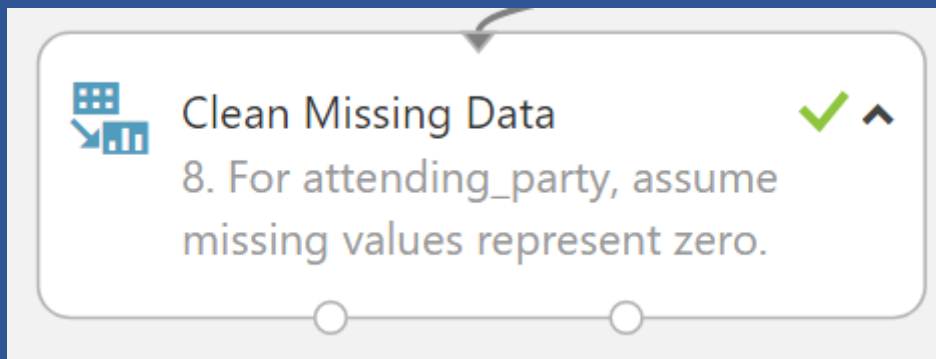
Replacement value

☒ Generate missing value indicator...

Missing value handling

Replace missing values with 0

- Drag & drop Clean Missing Data module
- Select column **attending_party**
- Configure “Cleaning mode” to **Custom substitution value**
- Replacement value = 0
- Comment = 8. For attending_party, assume missing values represent zero.
- Run/Visualize



Properties Project

Clean Missing Data

Columns to be cleaned

Selected columns:
Column names: attending_party

Launch column selector

Minimum missing value ratio

Maximum missing value ratio

Cleaning mode
Custom substitution value

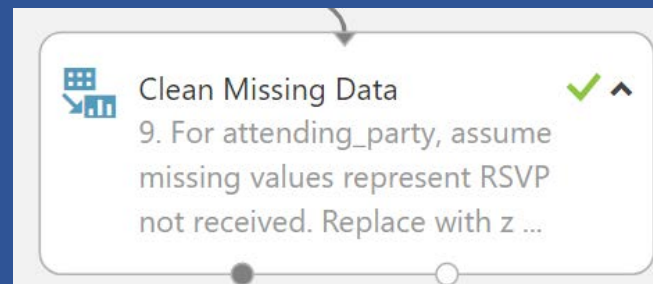
Replacement value

☐ Generate missing value indicato...

Missing value handling

Create an indicator variable for "missing."

- Drag & drop Clean Missing Data module
- Select column **attending_party**
- Configure "Cleaning mode" to **Custom substitution value**
- Check **Generate missing value indication column**
- Replacement value = 0
- Comment = 9. For attending_party, assume missing values represent RSVP not received. Replace with zero, but add a missing value indicator column.
- Run/Visualize



Properties Project

Clean Missing Data

Columns to be cleaned

Selected columns:
Column names: attending_party

Launch column selector

Minimum missing value ratio

Maximum missing value ratio

Cleaning mode
Custom substitution value

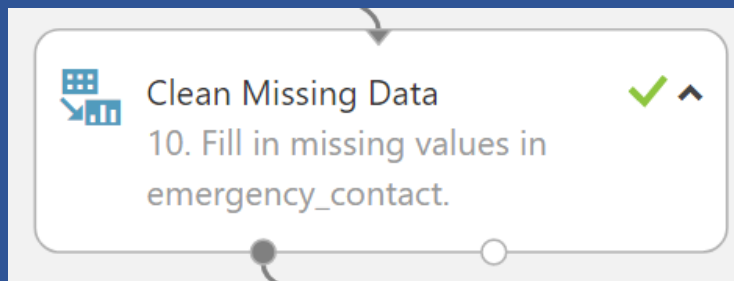
Replacement value

☒ Generate missing value indicator column

Missing value handling

Replace missing values with a string

- Drag & drop Clean Missing Data module
- Select column **emergency_contact**
- Configure “Cleaning mode” to **Custom substitution value**
- Replacement value = no
- Comment = 10. Fill in missing values in emergency_contact.
- Run/Visualize



Properties Project

Clean Missing Data

Columns to be cleaned

Selected columns:
Column names: emergency_contact

Launch column selector

Minimum missing value ratio

Maximum missing value ratio

Cleaning mode
Custom substitution value

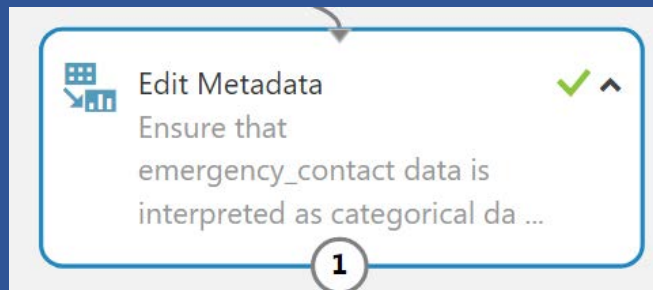
Replacement value

☐ Generate missing value indicator column

Missing value handling

Change metadata to categorical data

- Drag & drop **Edit Metadata**
- Select column **emergency_contact**
- Configure “Cleaning mode” to **Custom substitution value**
- Data type = String
- Categorical = Make categorical
- Comment = Ensure that emergency_contact data is interpreted as categorical data.
- Run/Visualize



Properties Project

Edit Metadata

Column

Selected columns:
Column names: emergency_contact

Launch column selector

Data type

String

Categorical

Make categorical

Fields

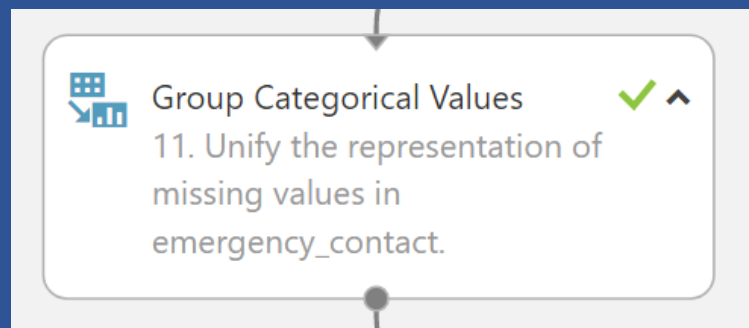
Unchanged

New column names

Missing value handling

Add an indicator variable showing which strings are considered "missing."

- Drag & drop **Group Categorical Values**
- Select column **emergency_contact**
- Configure "Cleaning mode" to **Custom substitution value**
- Output mode = Append
- Default level name = present
- New number of levels = 2
- Name of new level 1 = absent
- Comma-separate list of level to map to new level 1 = no,NA,n/a,None,_, "",empty,null
- Comment = 11. Unify the representation of missing values in emergency_contact.
- Run/Visualize



Properties Project

Group Categorical Values

Selected columns

Selected columns:
Column names: emergency_contact

Launch column selector

Output mode

Append

Default level name

present

New number of levels

2

Name of new level 1

absent

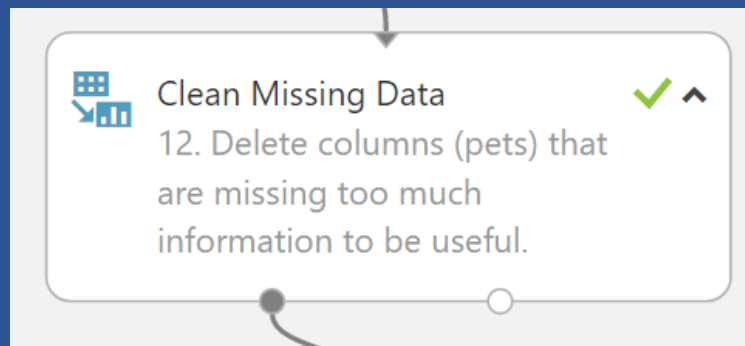
Comma-separated list of old levels to map..

no,NA,n/a,None,_, "",empty,null

Missing value handling

Delete columns that are missing too many values to be useful

- Drag & drop **Clean Missing Data**
- Select column **pets**
- Configure “Cleaning mode” to **Remove entire column**
- Comment = 12. Delete columns (pets) that are missing too much information to be useful.
- Run/Visualize



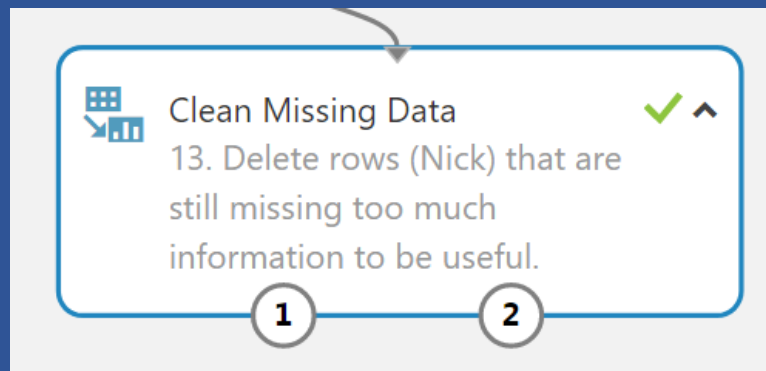
The screenshot shows the 'Properties' pane for the 'Clean Missing Data' module. The pane has a tab labeled 'Properties' and a sub-tab labeled 'Clean Missing Data'. The settings are as follows:

- Columns to be cleaned:** A box labeled 'Selected columns:' with 'Column names: pets' below it. A button 'Launch column selector' is below this box.
- Minimum missing value ratio:** A text input field containing '.5'.
- Maximum missing value ratio:** A text input field containing '1'.
- Cleaning mode:** A dropdown menu with 'Remove entire column' selected.

Missing value handling

Delete rows that are missing critical values

- Drag & drop **Clean Missing Data**
- Select column **entree.emergency_contact**
- Configure “Cleaning mode” to **Remove entire row**
- Comment = 13. Delete rows (Nick) that are still missing too much information to be useful.
- Run/Visualize



Properties Project

Clean Missing Data

Columns to be cleaned

Selected columns:
Column names:
entree,emergency_contact

Launch column selector









Minimum missing value ratio

Maximum missing value ratio

Cleaning mode
Remove entire row

Missing value handling

Final result

Column 0	age	years_seniority	income	parking_space	attending_party	entree	emergency_contact	parking_space_IsMissing
								
Tony	48	27	250	1	5	shrimp	Pepper	false
Donald	67	25	86	10	2	beef	Jane	false
Henry	69	21	95	6	1	chicken	Janet	false
Janet	62	21	110	3	1	beef	Henry	false
Bruce	37	14	63	-99	1	veggie	NA	true
Steve	83	12	77	7	1	chicken	n/a	false
Clint	27	9	118	9	0	shrimp	None	false
Wanda	19	7	52	2	2	shrimp	empty	false
Carol	46	3	127	11	1	veggie	""	false
Mandy	44	2	68	8	1	chicken	null	false

Missing value handling

More information

Clean Missing Data: Specifies how to handle the values missing from a dataset

<https://msdn.microsoft.com/library/azure/d2c5ca2f-7323-41a3-9b7e-da917c99f0c4>

This Experiment

<https://gallery.cortanaintelligence.com/Experiment/Missing-values>