



Data Preparation and Exploratory Data Analysis

Citizen Analytics – An Initiative by Data Science Team

START ►

© 2020 Petroliaam Nasional Berhad (PETRONAS)

All rights reserved. No part of this document may be reproduced in any form possible, stored in a retrieval system, transmitted and/or disseminated in any form or by any means (digital, mechanical, hard copy, recording or otherwise) without the permission of the copyright owner.

Learning Objectives

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



01

Understand the concept of Exploratory Data Analysis (EDA) and Statistics.

02

Identify ways to treat outliers and clean the data.

03

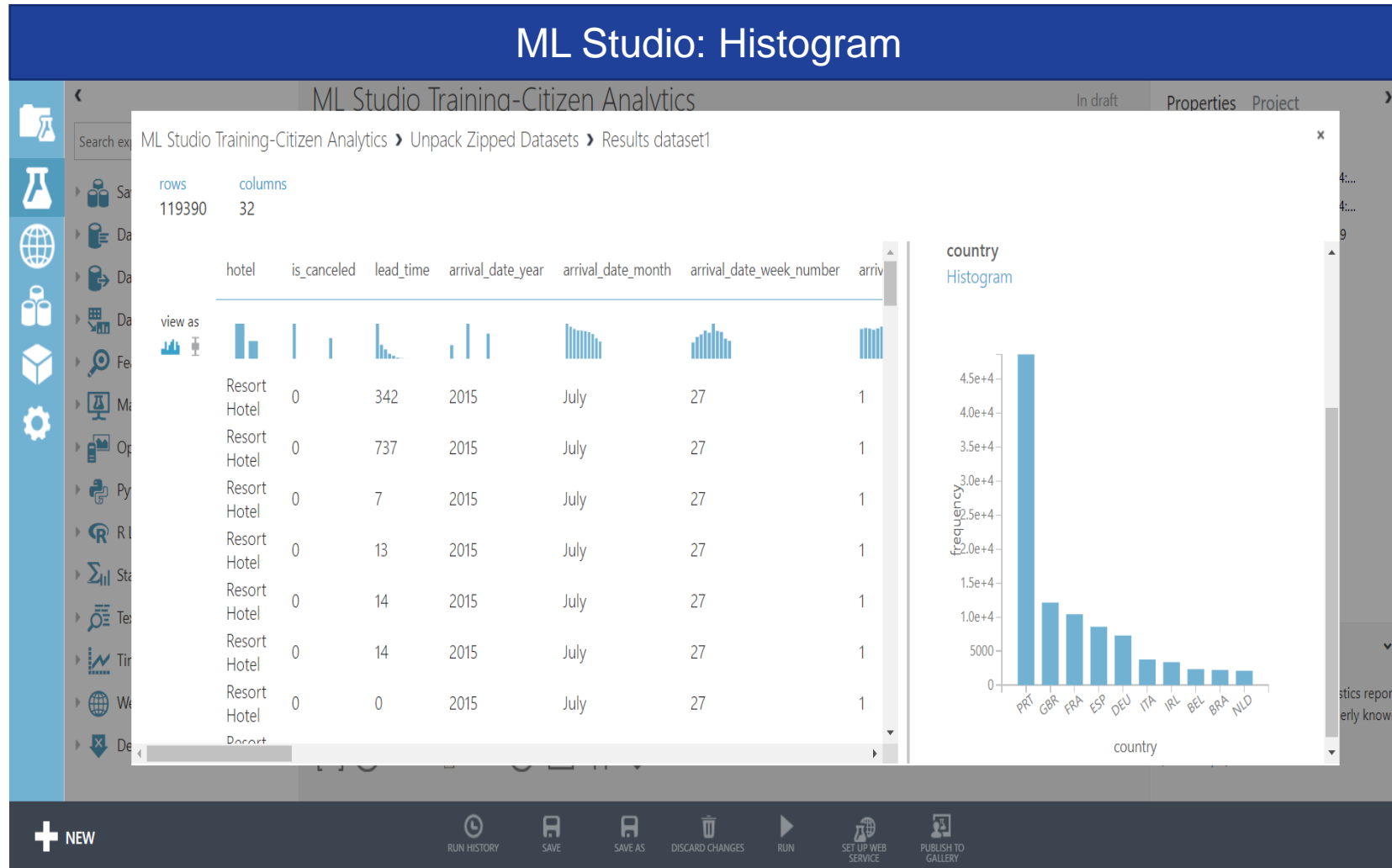
Identify ways to treat class imbalance, normalize data and joining multiple data.

Content

01. Exploratory Data Analysis	04	05. Imbalance Data	25
a. Graphical Techniques used in EDA		a. Class imbalance	
b. Distribution in Box-Plot and Histogram		06. Data Normalization	28
02. Statistics	08	a. Data Normalization and Example	
a. Statistics – Measures of Central Tendency		07. Joining Multiple Data	32
b. Statistics – Measures of Dispersion		a. Joining Data and Example	
03. Outlier Treatment	12	08. Summary	37
a. Outlier Treatments and Example			
04. Cleaning Missing Value	19		
a. Converting data types and columns' name			
b. Clean the dataset without NA values			
c. Mathematical Operation			

Exploratory Data Analysis

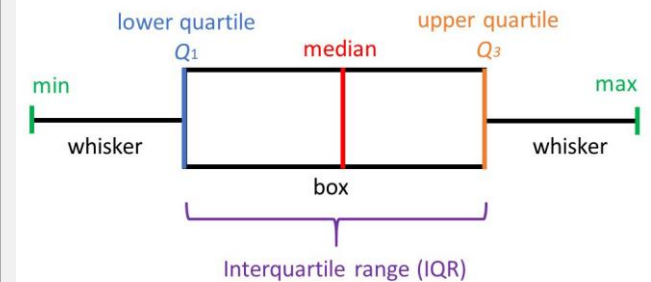
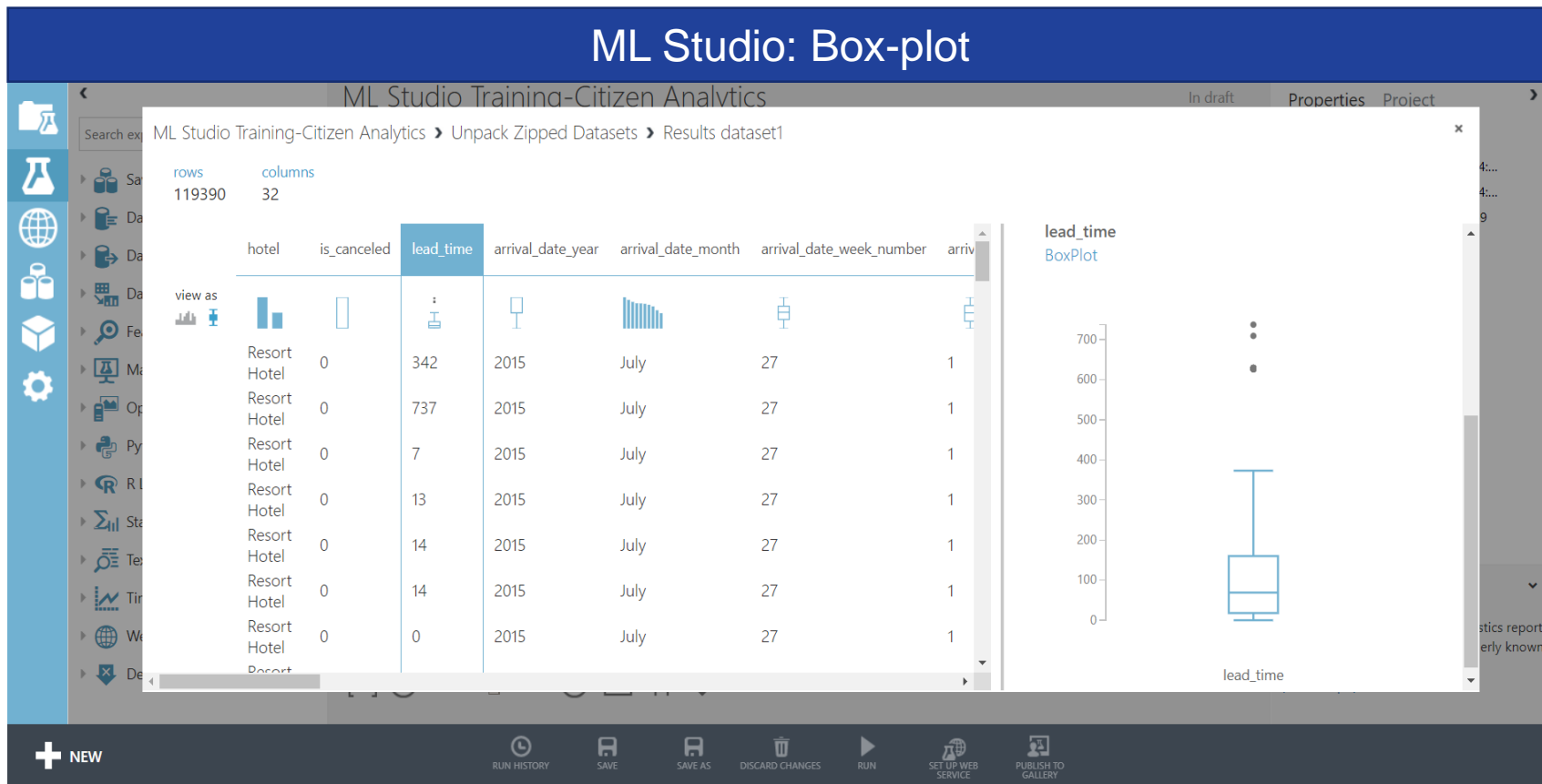
Graphical Techniques used in EDA (Histogram)



A histogram is a graphical display of data using **bars of different heights**, showing the shape and spread of continuous data. The taller the bar, the higher its occurrences.

Graphical Techniques used in EDA (Box-plot)

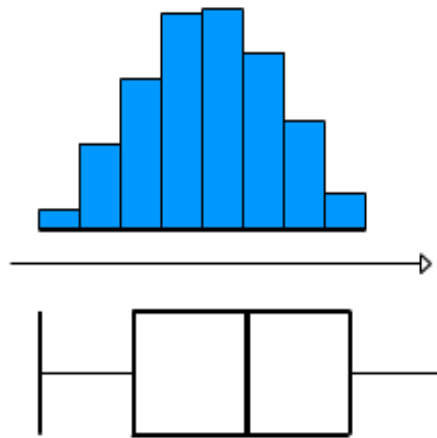
- A box plot (also known as box and whisker plot) is a chart that used to show the distribution of numerical data and its skewness by showing its quartiles and mean as well as its range.



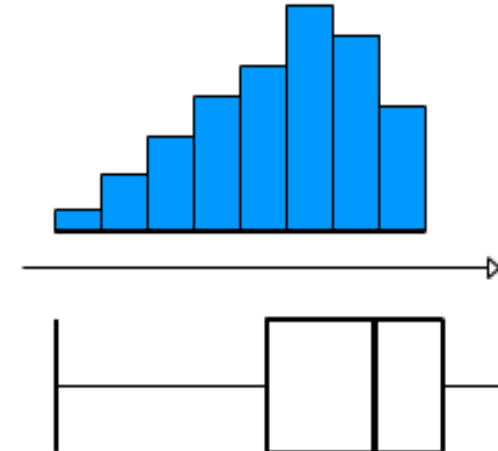
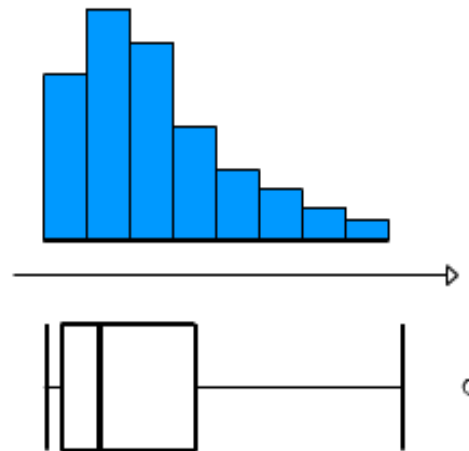
Distribution in Box-Plot and Histogram

- In symmetric distributions (normal), the mean, median, and mode are the same.
- In skewed data, the mean and median lie further toward the skew than the mode, either skewed to the right or left.

Normally distributed



Skewed

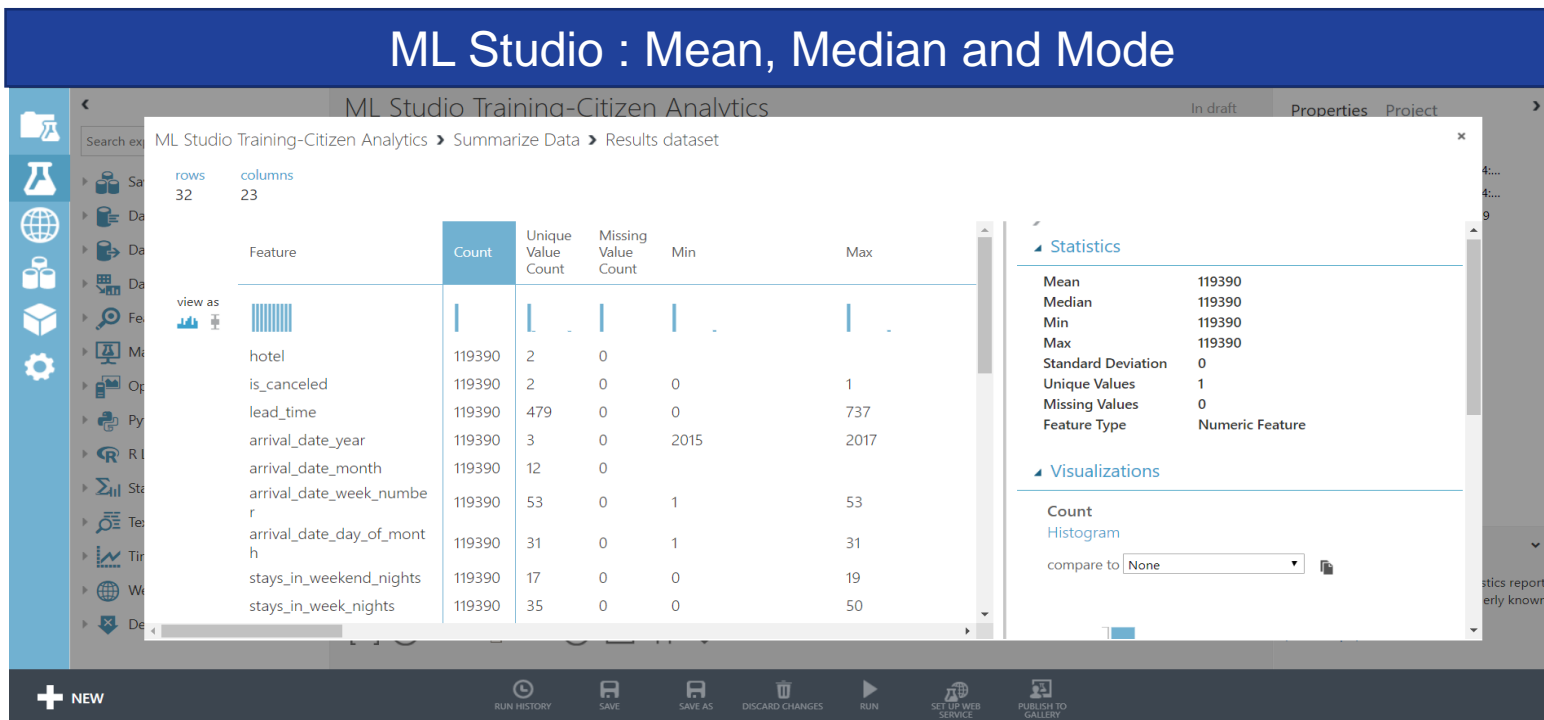


Statistics

Statistics – Measures of Central Tendency

A Measure of Central Tendency is a single value that attempts to describe the set of data by identifying the central position within that set of data.

- Mean
- Median
- Mode



Mean is the average value of a numeric dataset

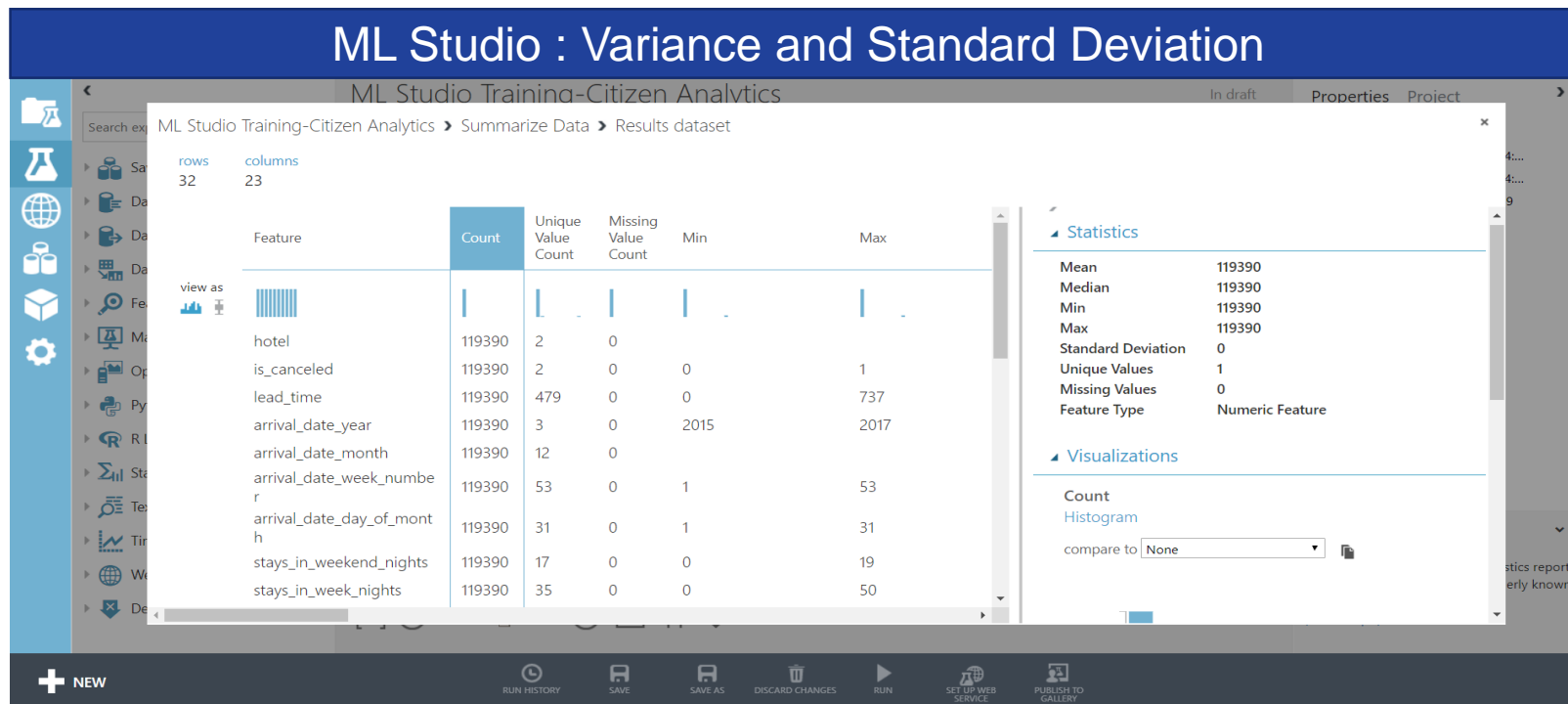
Mode is a dataset that has the highest frequency

Median is the centric value of ascending ordered-dataset

Statistics – Measures of Dispersion

Dispersion is a way of describing how spread out a set of data is.

- Range – Difference of maximum and minimum value
- Interquartile Range – Difference of first quartile and third quartile
- Variance
- Standard Deviation



Variance is the average squared distance between the mean and each data value.

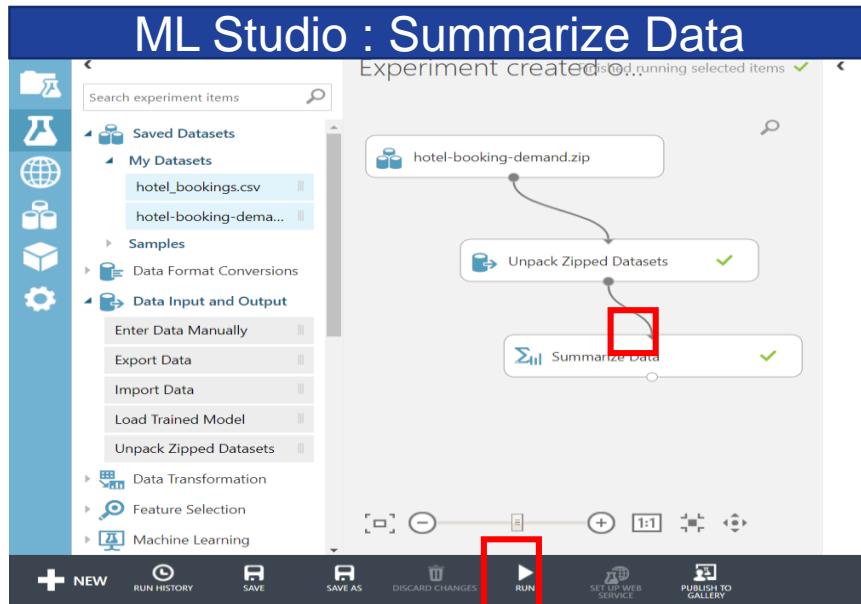
Standard deviation is just a square root of the variance, measuring how disperse is the data from the mean.



$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$$

Summary of the data

‘Summarize Data’ module allows us to see all related statistical measures such as mean, median, quartiles etc in just one pop out window.



Drag and drop “**Summarize Data**” module, click “**Run**” and later right-click on the circle to view Statistics Summary of your dataset of follows

The screenshot shows the 'Microsoft Azure Machine Learning Studio (classic)' interface. A pop-up window titled 'Experiment created on 5/2/2020 > Summarize Data > Results dataset' is displayed. It shows the dataset has 32 rows and 23 columns. A table of statistical measures is shown, with columns for Feature, Count, Unique Value Count, Missing Value Count, Min, Max, Mean, Mean Deviation, and 1st Quartile. The features listed are hotel, is_canceled, lead_time, arrival_date_year, arrival_date_month, arrival_date_week_number, arrival_date_day_of_month, stays_in_weekend_nights, and stays_in_week_nights.

Feature	Count	Unique Value Count	Missing Value Count	Min	Max	Mean	Mean Deviation	1st Quartile
hotel	119390	2	0					
is_canceled	119390	2	0	0	1	0.370416	0.466416	0
lead_time	119390	479	0	0	737	104.011416	84.671975	18
arrival_date_year	119390	3	0	2015	2017	2016.156554	0.574877	2016
arrival_date_month	119390	12	0					
arrival_date_week_number	119390	53	0	1	53	27.165173	11.549925	16
arrival_date_day_of_month	119390	31	0	1	31	15.798241	7.578563	8
stays_in_weekend_nights	119390	17	0	0	19	0.927599	0.807995	0
stays_in_week_nights	119390	35	0	0	50	2.500302	1.364287	1

Outlier Treatment

Outliers treatment

Outliers are observations that is far away from other observations and might create bias if it is not treated well

ML Studio : Clip Values

ML Studio Training-Citizen Analytics

In draft
Draft saved at 11:57:47 AM

Look for “Clip Values” and drag and drop it to the middle.

1

2

3

On “**Set of thresholds**”, there are 3 choices which are “**ClipPeaks**”, “**ClipSubpeaks**” and “**ClipPeaksAndSubpeaks**”:

1

“**ClipPeaks**” is clipping the higher threshold. Choose this if your variable has more outliers at the upper level of the value of the variable

2

“**ClipSubpeaks**” is clipping the lower threshold. Choose this if your variable has more outliers at the lower level of the value of the variable

3

“**ClipPeaksAndSubpeaks**” is clipping threshold at the upper and lower level. Choose this if your variable has outliers both at the upper and lower level of the value of the variable

Outliers treatment

Outliers are observations that is far away from other observations and might create bias if it is not treated well

The screenshot displays the ML Studio 'Clip Values' module. On the left, a workflow diagram shows the sequence of operations: 'hotel-booking-demand.zip' is unpacked, then metadata is edited, followed by cleaning missing data, applying math operations, and finally using the 'Clip Values' module (highlighted with a red box and the number 1). On the right, the 'Properties' panel for the 'Clip Values' module is shown. It includes a 'Set of thresholds' dropdown set to 'ClipPeaks', an 'Upper threshold' dropdown set to 'Constant' with a value of '1', and an 'Upper substitute value' dropdown set to 'Threshold'. The 'List of columns' section shows 'Selected columns: Column type: Numeric, All'. A 'Launch column selector' button is also visible. The bottom of the interface shows a toolbar with icons for 'RUN HISTORY', 'SAVE', 'SAVE AS', 'DISCARD CHANGES', 'RUN', 'SET UP WEB SERVICE', and 'PUBLISH TO GALLERY'.

ML Studio : Clip Values

ML Studio Training-Citizen Analytics

In draft

Draft saved at 11:57:47 AM

Properties Project

Clip Values

Set of thresholds

ClipPeaks

Upper threshold

Constant

Constant value for upp...

1

Upper substitute value

Threshold

List of columns

Selected columns:

Column type: Numeric, All

Launch column selector

Overwrite flag

Quick Help

Detects outliers and clips or replaces their values (more help...)

Look for “Clip Values” and drag and drop it to the middle.

“**ClipPeaks**” is clipping higher threshold. Choose this if your variable has more outliers at the upper level of the value of the variable

At “**Upper threshold**”, choose “**Constant**” if you want to clip all the outliers starting from a certain number. choose “**Percentile**” if you want to clip all the outliers to a certain percentage.

At Upper substitute value’, you can choose either “**Threshold**”, “**Median**”, “**Mean**” and “**Missing**” to replace the value of the outliers.

Lastly, select which column you want to treat for its outliers.

Outliers treatment - Example

Dataset = Adult Census Income Binary

Set of thresholds (ClipPeaks) > Upper threshold (Constant) > Constant value for upper threshold (70) > Upper substitute value (Threshold) > Column (Age)

ML Studio : Clip Values

Experiment created... Finished running ✓

Properties Project

Clip Values

Set of thresholds
ClipPeaks

Upper threshold
Constant

Constant value for upp...
70

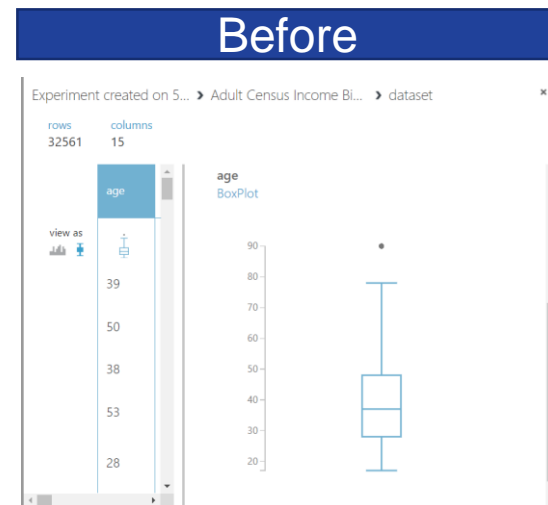
Upper substitute value
Threshold

List of columns
Selected columns:
Column names: age

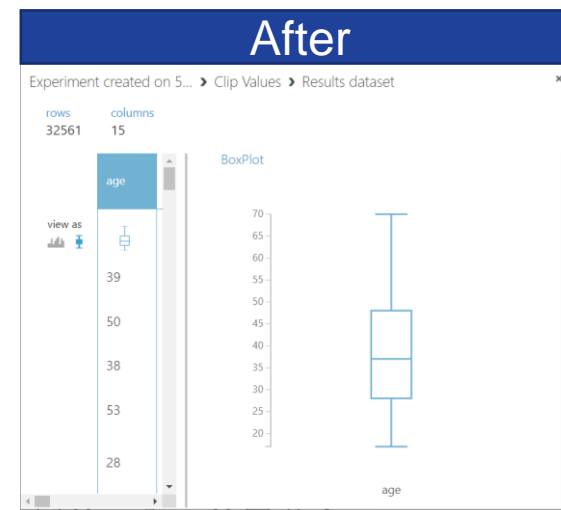
Launch column selector

☒ Overwrite flag

Quick Help
Detects outliers and clips or replaces their values
(more help...)



The dot represents the outliers in this data



The data is already clipped and the outliers is no longer there as the value is replaced by using the new upper threshold.

Outliers treatment - Example

Dataset = Adult Census Income Binary

Set of thresholds (ClipSubpeaks) > Lower threshold (Percentile) > Percentile value for upper threshold (5) > Upper substitute value (Threshold) > Column (education-num)

ML Studio : Clip Values

Experiment created... In draft
Draft saved at 12:53:23 PM

Adult Census Income Binary...

Clip Values 1

Properties Project

Clip Values

Set of thresholds
ClipSubpeaks

Lower threshold
Percentile

Percentile number for l...
5

Lower substitute value
Threshold

List of columns
Selected columns:
Column names:
education-num

Launch column selector

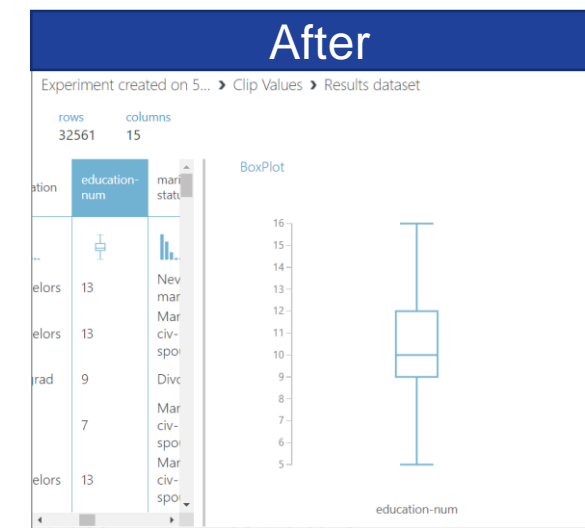
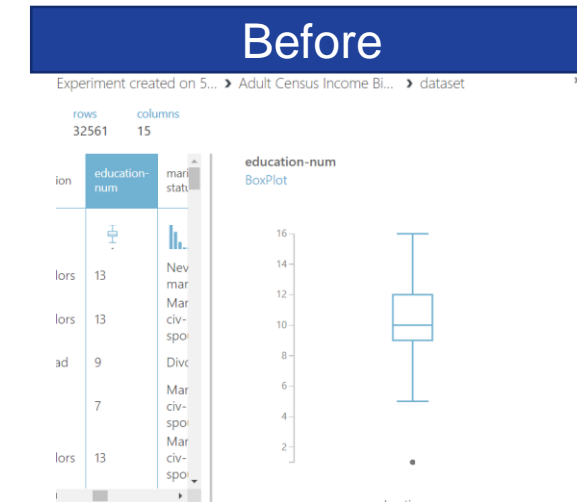
☒ Overwrite flag

☐ Add indicator colu...

START TIME 5/16/20...
END TIME 5/16/20...
ELAPSED TIME 0:00:02.7

Quick Help

+ NEW RUN HISTORY SAVE SAVE AS DISCARD CHANGES RUN SET UP WEB SERVICE PUBLISH TO GALLERY



Outliers treatment - Example

Dataset = Adult Census Income Binary

Set of thresholds (ClipPeaksAndSubpeaks) > Threshold (Percentile) > Percentile value for upper threshold (89) > Percentile value for lower threshold (17) > Upper substitute value (Threshold) > Column (hours-per-week)

ML Studio : Clip Values

Experiment created... In draft
Draft saved at 1:31:28 PM

Adult Census Income Binary...

Clip Values 1

Properties Project

Clip Values

Set of thresholds
ClipPeaksAndSubpeaks

Threshold
Percentile

Percentile number of u...
89

Percentile number of lo...
17

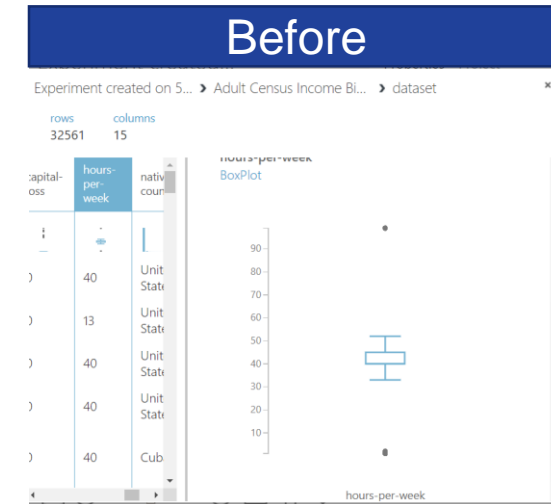
Substitute value for pe...
Threshold

Substitute value for sub...
Threshold

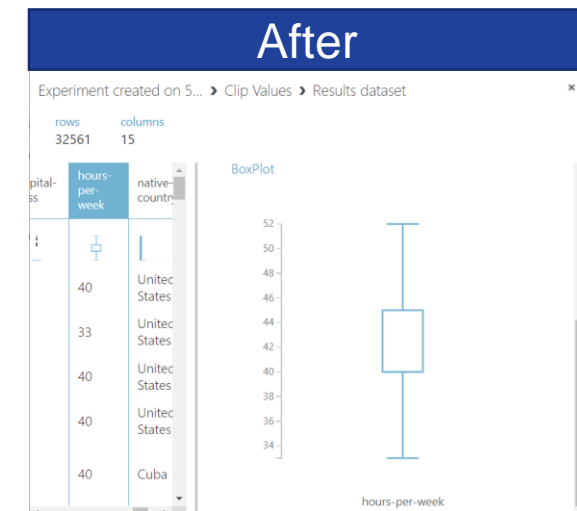
List of columns
Selected columns:
Column names: hours-per-week
Launch column selector

Quick Help

NEW RUN HISTORY SAVE SAVE AS DISCARD CHANGES RUN SET UP WEB SERVICE PUBLISH TO GALLERY



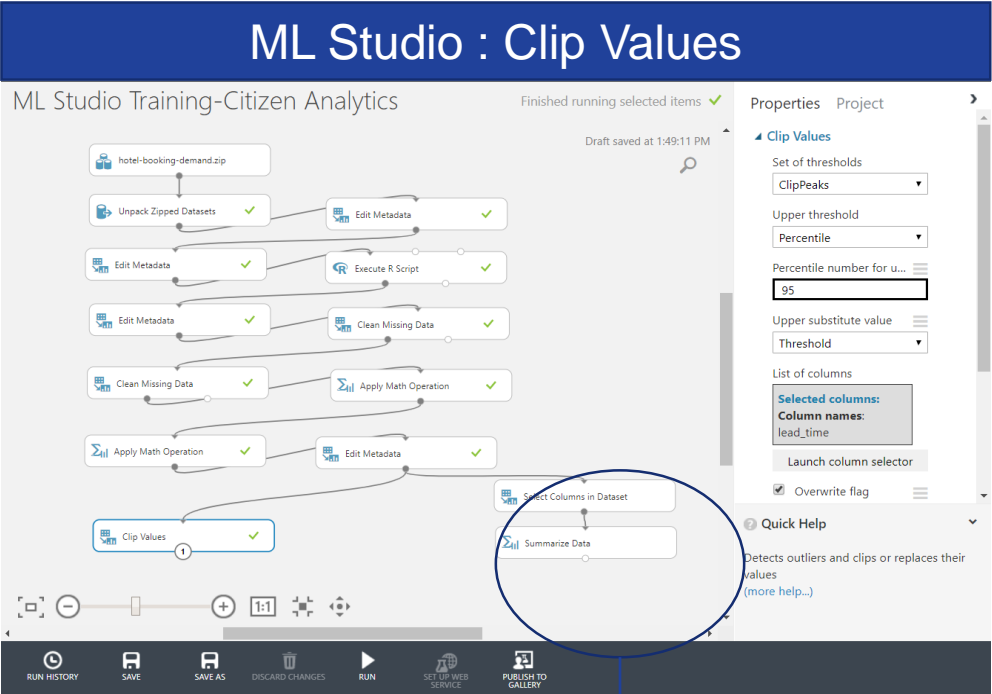
The dot represents the outliers in this data



The data is already clipped and the outliers is no longer there as the value is replaced by using the new upper and lower percentile of threshold.

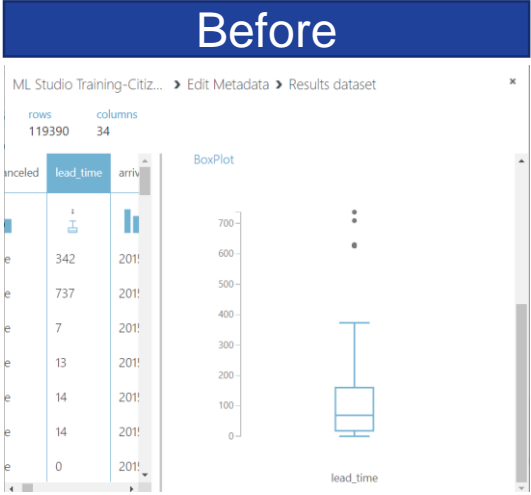
Outliers treatment

Set of thresholds (ClipPeaks) > Upper threshold (Percentile) > Constant value for upper threshold (95) > Upper substitute value (Threshold) > Column (lead_time)

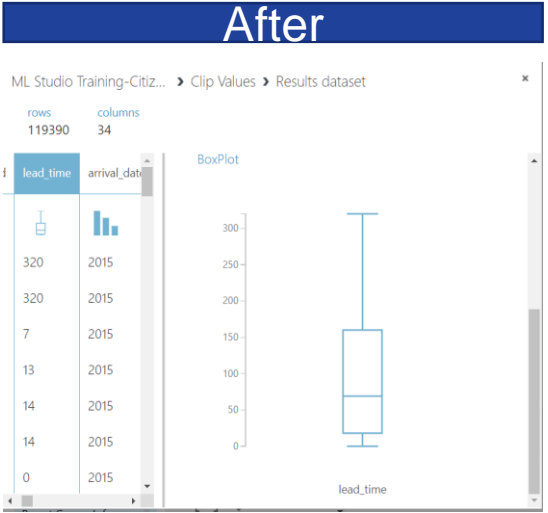


Result

	rows	columns
	1	23
	Sample Skewness	Sample Kurtosis
	1.34655	1.696449
	P0.5	P1
	0	0
	P5	P95
	0	320
	P99	P99.5
	444	476.11



The dot represents the outliers in this data



The data is already clipped and the outliers is no longer there as the value is replaced by using the new upper threshold.

Cleaning Missing Value

Converting data types and columns' name

ML Studio : Edit Metadata

ML Studio Training-Citizen Analytics

Drag and drop
"Edit Metadata"
module

hotel-booking-demand.zip

Unpack Zipped Datasets

Edit Metadata

Properties Project

Edit Metadata

Column

Selected columns:
Column names:
is_canceled, arrival_date_year

Launch column selector

Data type
Unchanged

Categorical
Make categorical

Fields
Unchanged

New column names

Quick Help

Edits metadata associated with columns in a dataset. Formerly known as Metadata Editor.

Select columns

BY NAME
WITH RULES

AVAILABLE COLUMNS

All Types search columns

hotel
lead_time
arrival_date_week_number
arrival_date_day_of_month
stays_in_weekend_nights
stays_in_week_nights
adults
children
babies
previous_bookings_not_canceled
booking_changes
days_in_waiting_list
adr
required_car_parking_spaces
total_of_special_requests
reservation_status_date

16 columns available

SELECTED COLUMNS

All Types search columns

is_canceled
arrival_date_year
arrival_date_month
meal
country
market_segment
distribution_channel
is_repeated_guest
previous_cancellations
reserved_room_type
assigned_room_type
deposit_type
agent
company
customer_type
reservation_status

16 columns selected

Click **"Launch column selector"** to select columns that you want to change either its datatype, data category or its name

Change these variables to Categorical data by selecting 'Make Categorical' in 'Categorical' section.

Converting data types and columns' name

The screenshot displays the Microsoft Azure Machine Learning Studio (classic) interface. The main workspace shows a workflow starting with 'hotel-booking-demand.zip', followed by 'Unpack Zipped Datasets', and then two 'Edit Metadata' modules. The second 'Edit Metadata' module is highlighted with a red box and the number '1'. A text overlay says 'Copy and paste "Edit Metadata" module'. The right sidebar shows the 'Properties' pane for the selected 'Edit Metadata' module, with the 'Edit Metadata' tab active. The 'Column' section shows 'Selected columns: is_canceled, is_repeated, g...'. Below this, the 'Launch column selector' button is visible. The 'Data type' is set to 'Boolean', and the 'Categorical' section is set to 'Make categorical'. The 'Fields' section is set to 'Unchanged'. The 'New column names' section is empty. The 'Select columns' dialog is open in the foreground, showing a list of available columns on the left and a list of selected columns on the right. The 'Available columns' list includes 'hotel', 'lead_time', 'arrival_date_year', 'arrival_date_month', 'arrival_date_week_number', 'arrival_date_day_of_month', 'stays_in_weekend_nights', 'stays_in_week_nights', 'adults', 'children', 'babies', 'meal', 'country', 'market_segment', 'distribution_channel', and 'previous_bookings'. The 'Selected columns' list includes 'is_canceled' and 'is_repeated_guest'. The dialog has a 'BY NAME' tab and a 'WITH RULES' tab. The 'WITH RULES' tab is selected. The 'Available columns' list is filtered by 'All Types'. The 'Selected columns' list is filtered by 'All Types'. The dialog has a search bar and a '30 columns available' indicator. The '2 columns selected' indicator is also present.

ML Studio : Edit Metadata

Microsoft Azure Machine Learning Studio (classic)

ML Studio Training-Citizen Analytics

In draft

Draft saved at 1:03:33 PM

Copy and paste "Edit Metadata" module

Click "Launch column selector" to select columns that you want to change either its datatype, data category or its name

Select columns

BY NAME

WITH RULES

AVAILABLE COLUMNS

All Types search columns

hotel
lead_time
arrival_date_year
arrival_date_month
arrival_date_week_number
arrival_date_day_of_month
stays_in_weekend_nights
stays_in_week_nights
adults
children
babies
meal
country
market_segment
distribution_channel
previous_bookings

30 columns available

SELECTED COLUMNS

All Types search columns

is_canceled
is_repeated_guest

2 columns selected

Change these variables to Categorical data by selecting 'Make Categorical' in 'Categorical' section and 'Boolean' for datatype.

Clean the dataset without NA values

ML Studio : Execute R script

ML Studio Training-Citizen Analytics

In draft

Draft saved at 1:48:44 PM

Properties Project

Execute R Script

R Script

```
1 # Map 1-based optional input ports to variables
2 dataset1 <- maml.mapInputPort(1); # class: data.frame
3 dataset1$country[dataset1$country == 'NULL'] = NA;
4 dataset1$agent[dataset1$agent == 'NULL'] = NA;
5 dataset1$company[dataset1$company == 'NULL'] = NA;
6 dataset1$children[dataset1$children == 'NA'] = NA;
```

Random Seed

R Version

Microsoft R Open 3.2.2

START TIME 5/13/2020 1:46:52 PM

END TIME 5/13/2020 1:47:02 PM

ELAPSED TIME 0:00:10.246

STATUS CODE Finished

STATUS DETAILS None

[View output log](#)

Quick Help

hotel-booking-demand.zip

Unpack Zipped Datasets

Edit Metadata

Edit Metadata

Execute R Script

Edit Metadata

Drag and drop "Execute R Script" module

1 2

RUN HISTORY SAVE SAVE AS DISCARD CHANGES RUN SET UP WEB SERVICE PUBLISH TO GALLERY

Change "NULL" and "NA" values in 'Children', 'Agent', 'Company' and 'Country' to NA

Copy and paste this R code in "R Script" space.

```
# Map 1-based optional input ports to variables
dataset1 <- maml.mapInputPort(1); # class: data.frame
dataset1$country[dataset1$country == 'NULL'] = NA;
dataset1$agent[dataset1$agent == 'NULL'] = NA;
dataset1$company[dataset1$company == 'NULL'] = NA;
dataset1$children[dataset1$children == 'NA'] = NA;
maml.mapOutputPort("dataset1");
```

Clean the dataset with NA values

ML Studio : Clean Missing Value

Data with missing values

Finished running selected items ✓

Draft saved at 2:30:58 PM

hotel_bookings.csv

Summarize Data

Clean Missing Data

Clean Missing Data

Properties Project

Clean Missing Data

Columns to be cleaned

Selected columns:
Column names: children, agent, company

Launch column selector

Minimum missing value ratio
0

Maximum missing value ratio
1

Cleaning mode
Custom substitution value

Replacement value
0

☐ Generate missing value indicator column

START TIME 5/12/2020 2:30:12 PM

Quick Help

Specifies how to handle the values missing from a dataset
(more help...)

RUN HISTORY SAVE SAVE AS DISCARD CHANGES RUN SET UP WEB SERVICE PUBLISH TO GALLERY

Replace NA values in
'Children', 'Agent', 'Company'
to 0 while 'Country' to
Unknown

Mathematical Operation

ML Studio : Apply Math Operation

ML Studio Training-Citizen Analytics

Finished running ✓

Drag and drop “**Apply Math Operation**” if you wish to apply mathematical formula on the columns with integer datatype

Properties Project

Apply Math Operation

Category: Operations

Basic operation: Add

Operation argument type: ColumnSet

Operation argument:

Selected columns:
Column names: adults

Launch column selector

Column set:
Selected columns:
Column names: children

Launch column selector

Quick Help
Applies a mathematical operation to column values
(more help...)

RUN HISTORY SAVE SAVE AS DISCARD CHANGES RUN SET UP WEB SERVICE PUBLISH TO GALLERY

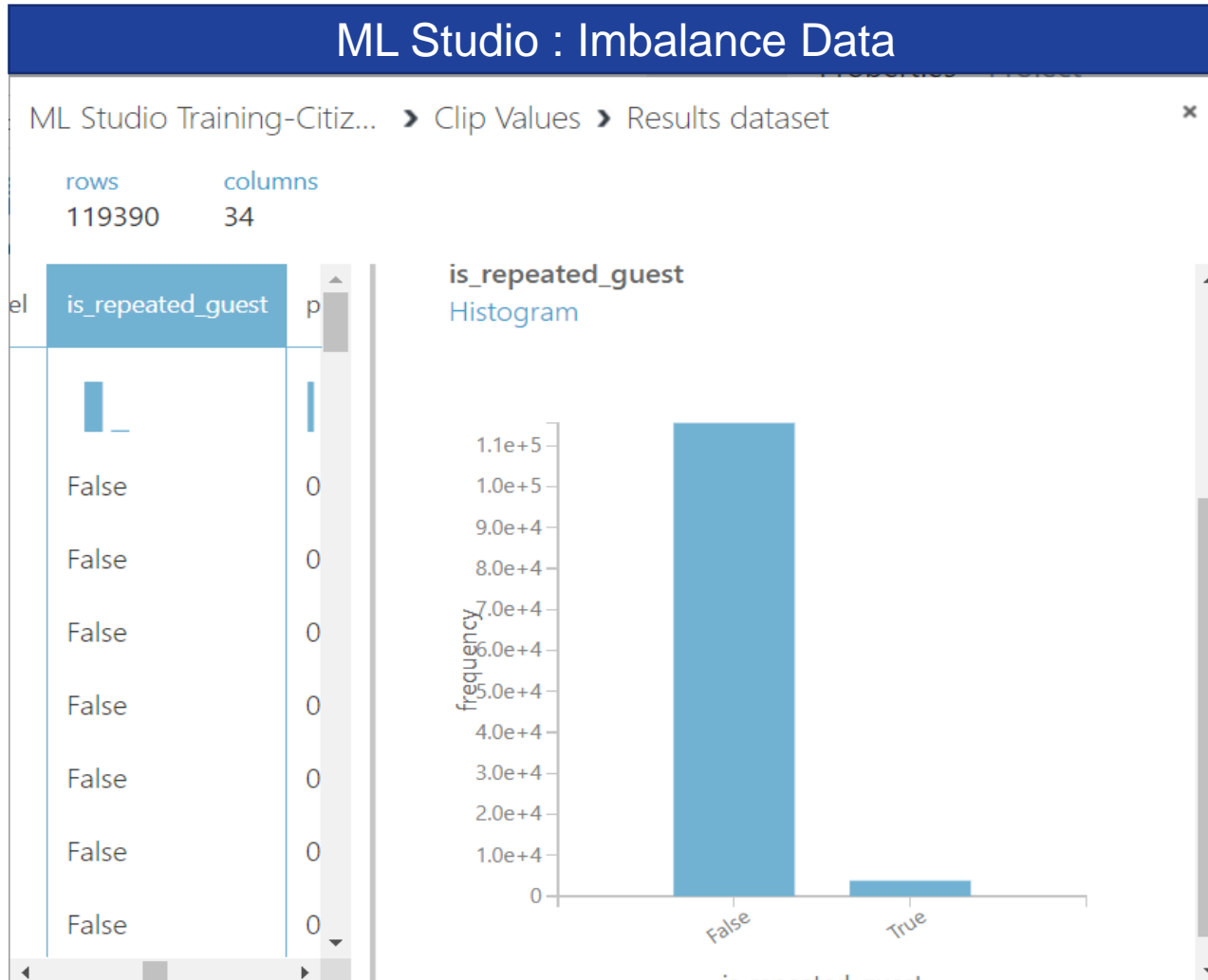
On “**Category**”, you can select what kind of mathematical function you want to apply to integer columns

Add values in columns ‘Adult’, ‘Children’ and then after getting the result, add to ‘Babies’ and rename the new columns

Imbalance Data

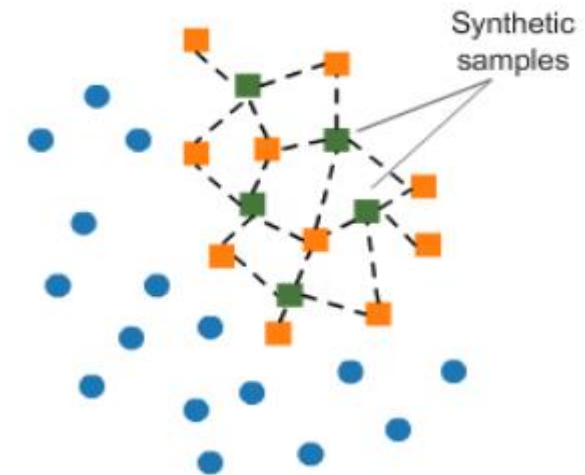
Class Imbalance

Imbalanced data happen when the classes are not represented equally



Dealing with imbalanced dataset using SMOTE – Synthetic Minority Oversampling Technique

SMOTE is just creating new synthetic datapoints between two datapoints



Class Imbalance

Dealing with imbalance data by oversampling method

ML Studio : Imbalance Data

ML Studio Training-Citizen Analytics

In draft

Draft saved at 2:12:59 PM

Properties Project

SMOTE

Label column

Selected columns:
Column names:
is_repeated_guest

Launch column selector

SMOTE percentage
100

Number of nearest neighbors
1

Random seed
123

Quick Help

Increases the number of low incidence examples in a dataset using synthetic minority oversampling
[\(more help...\)](#)

At “**SMOTE Percentage**”, choose 100 if you want to produce the same number of observations from the imbalanced class.

“**Number of nearest neighbors**” is to set how many data points that you want to create between neighbors

Data Normalization

Data Normalization

Concept

- A method to standardize the range of independent variables value, where mostly the resulting range will be fitted within 0 to 1.

Why need to normalize?

- To make data more stable and the coefficient that will be derived is more reliable as the data is on the same scale.

$$z = \frac{x - \bar{x}}{\sigma}$$

Z-score Normalization

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Min-Max Normalization

$$z = \frac{1}{1 + \exp(x)}$$

Logistic Normalization

$$z = \text{lognormal.CDF}(x, \mu, \sigma)$$

LogNormal Normalization

$$= \frac{p(k|x; \theta)}{[E(Y|x)]^k e^{E(Y|x)}} \\ = \frac{1}{k!}$$

TanH Normalization

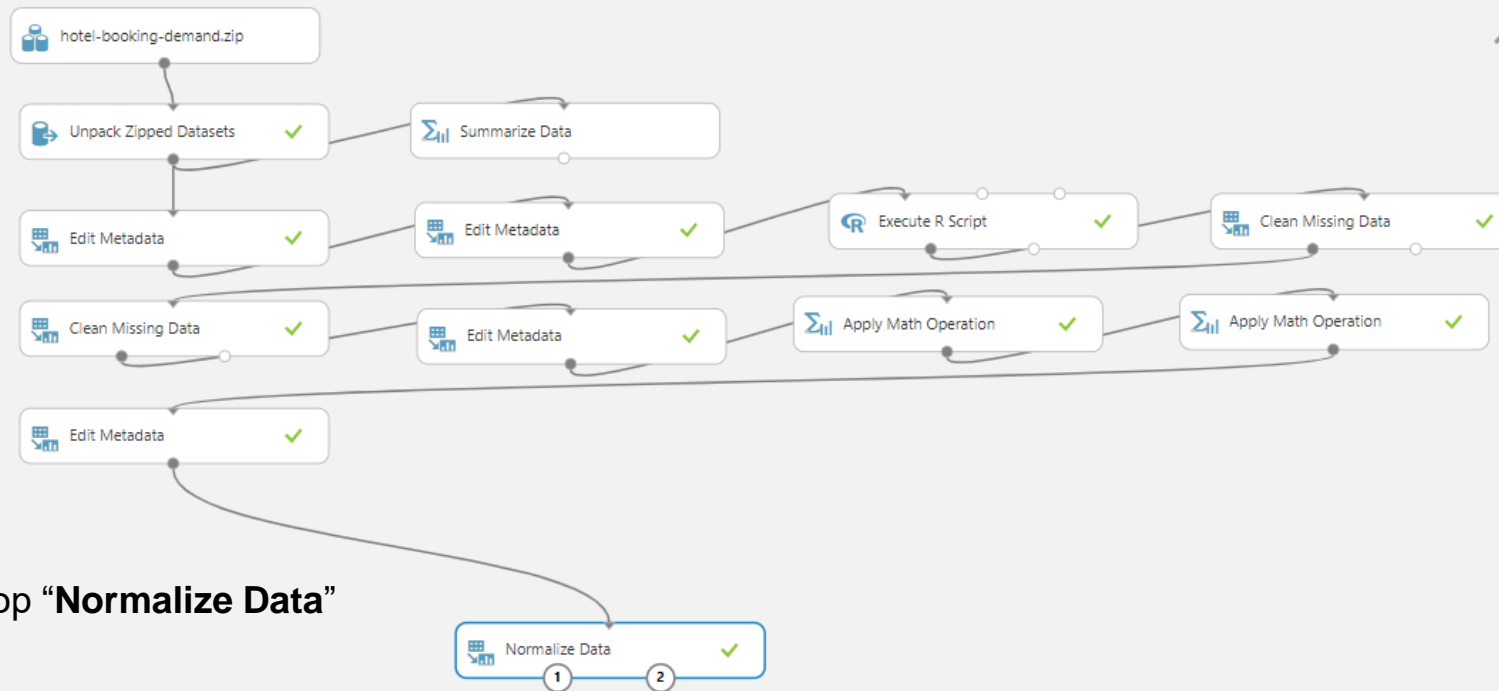
Data Normalization - Example

ML Studio : Normalize Data

ML Studio Data Import

Finished running selected items ✓

Properties Project



Drag and drop “**Normalize Data**”

Normalize Data

Transformation method

ZScore

☒ Use 0 for constant columns when c...

Columns to transform

Selected columns:

Column type: Numeric, All

Launch column selector

START TIME 10/26/2020 9:25:23 AM

END TIME 10/26/2020 9:25:26 AM

ELAPSED TIME 0:00:03.396

STATUS CODE Finished

STATUS DETAILS None

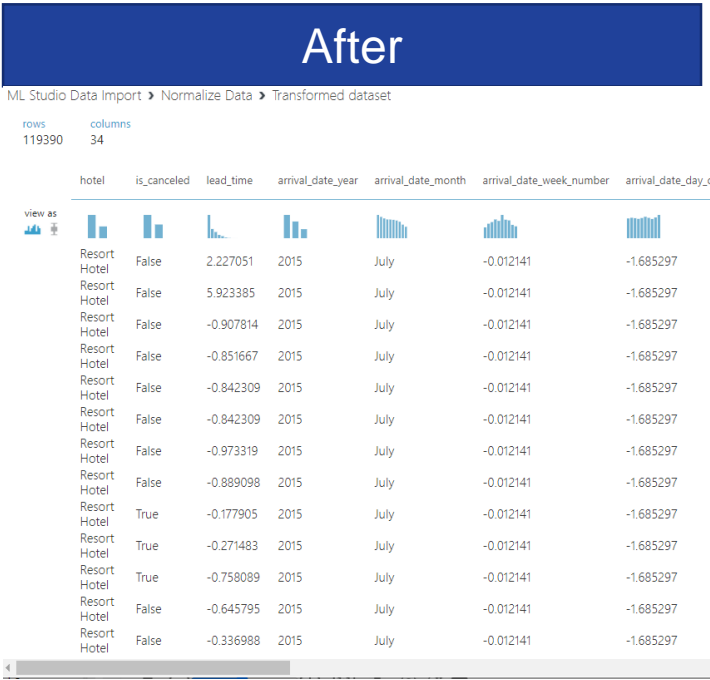
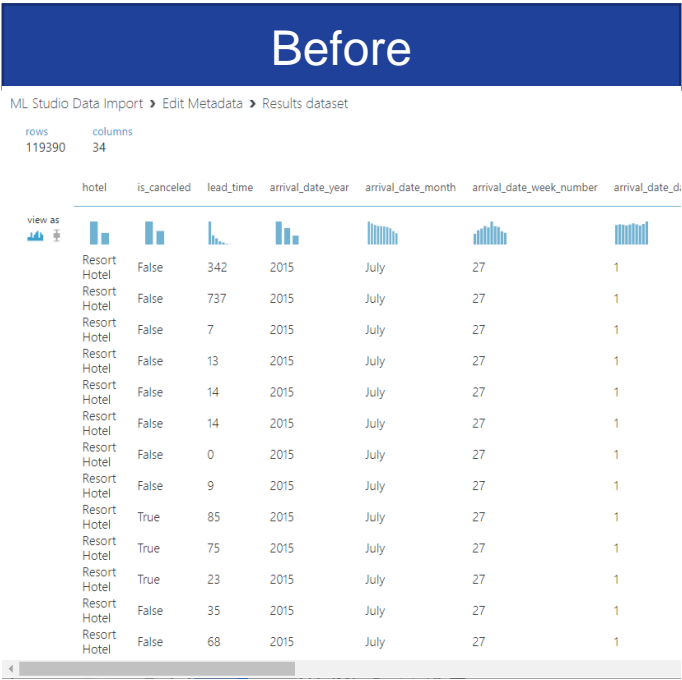
[View output log](#)

Web Service Parameters

Dataset to Unpack

At “**Transformation method**”, select type of normalization and select columns

Data Normalization - Example



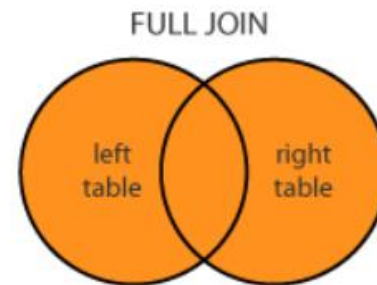
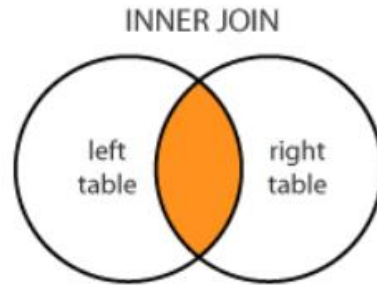
The original data with values ranges in hundred.

The normalized data with values ranges in tenth using Z score.

Joining Multiple Data

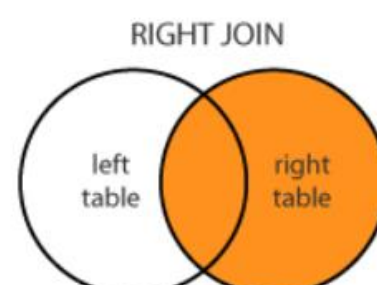
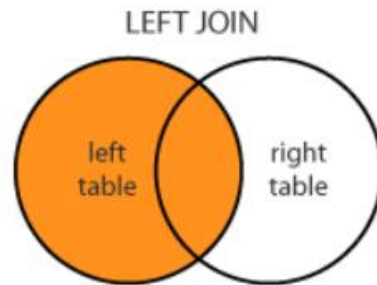
Joining data

Data is merged on the common column on both tables



Data is combined regardless of the columns name.

Data is merged on the common column but returns only the values from the left table while the key columns match.


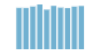








Data is merged on the common column but returns only the values from the right table while the key columns match.

Notes:

When a row in the left (right) table has no matching rows in the right (left) table, the returned row contains missing values for all columns that come from the right (left) table unless you specify a replacement value for missing values.

Joining data - Example

Movie Ratings Dataset				
Experiment created on 5/18/2020 ▶ Movie Ratings ▶ dataset				
rows 227472	columns 4			
	Userld	Movield	Rating	Timestamp
view as 				
	1	68646	10	1381620027
	1	113277	10	1379466669
	2	454876	8	1394818630
	2	790636	7	1389963947
	2	816711	8	1379963769
	2	1091191	7	1391173869
	2	1322269	7	1391529691
	2	1433811	8	1380453043
	2	1454468	8	1387016442
	2	1535109	8	1386350135
	2	1675434	8	1396688981
	2	1798709	10	1389948338

IMDB Movie Titles Dataset	
Experiment created on 5/18/2020 ▶ IMDB Movie Titles ▶ dataset	
rows 16614	columns 2
	Movie ID Movie Name
view as 	 
	8 Edison Kinetoscopic Record of a Sneeze (1894)
	91 Le manoir du diable (1896)
	417 Le voyage dans la lune (1902)
	628 The s of Dollie (1908)
	833 The Country Doctor (1909)
	1223 Frankenstein (1910)
	1740 The Lonedale Operator (1911)
	2101 Cleopatra (1912)
	2130 Linferno (1911)
	Fantmas - l'ombre de la

Both are original data from two different tables, aims to join this two tables on 'Movield' column.

Joining data - Example

The screenshot displays the ML Studio 'Joining Data' interface. At the top, a blue header reads 'ML Studio : Joining Data'. Below it, a status bar indicates 'Experiment created on 5/18/2020' and 'Finished running' with a green checkmark. The main workspace shows a workflow with two input nodes, 'Movie Ratings' and 'IMDB Movie Titles', connected to a 'Join Data' node. A green checkmark is next to the 'Join Data' node, and a circled '1' is next to its output. A text overlay says 'Drag and drop those data and "Join Data"'. On the right, the 'Properties' panel for the 'Join Data' node is visible. It shows 'Join key columns for L' with 'Selected columns: Column names: MovieId' and 'Join key columns for R' with 'Selected columns: Column names: Movie ID'. The 'Join type' is set to 'Inner Join'. At the bottom, a toolbar contains icons for 'RUN HISTORY', 'SAVE', 'SAVE AS', 'DISCARD CHANGES', 'RUN', 'SET UP WEB SERVICE', and 'PUBLISH TO GALLERY'.

Experiment created on 5/18/2020

Finished running ✓

Draft saved at 5:41:53 PM

Movie Ratings

IMDB Movie Titles

Join Data

1

Drag and drop those data and "Join Data"

Properties Project

Join Data

Join key columns for L

Selected columns: Column names: MovieId

Launch column selector

Join key columns for R

Selected columns: Column names: Movie ID

Launch column selector

Match case

Join type

Inner Join

Keep right key colu...

Quick Help

Joins two datasets on selected key columns. (more help...)

RUN HISTORY SAVE SAVE AS DISCARD CHANGES RUN SET UP WEB SERVICE PUBLISH TO GALLERY

Select which column has the same data attribute

At "Join Type", select either "Inner Join", "Left Outer Join", "Left Semi-Join" or "Full Outer Join"


- Left Outer Join = Left join
- Full Outer Join = Outer join
- For each of the rows in the left table that has no matching rows in the right table, the join results include a row containing missing values from the right table.
- For each of the rows in the right table that has no matching rows in the left table, the join results include a row containing missing values for all columns from the left table.
- Left Semi-Join = Left join but returns only the values when key columns match







Joining data - Example

ML Studio : Joining Data- Result

rows
227472

columns
6

view as 

	UserId	MovieId	Rating	Timestamp	Movie ID	Movie Name
						
	1	68646	10	1381620027	68646	The Godfather (1972)
	1	113277	10	1379466669	113277	Heat (1995)
	2	454876	8	1394818630	454876	Life of Pi (2012)
	2	790636	7	1389963947	790636	Dallas Buyers Club (2013)
	2	816711	8	1379963769	816711	World Z (2013)
	2	1091191	7	1391173869	1091191	Lone Survivor (2013)
	2	1322269	7	1391529691	1322269	August Osage County (2013)
	2	1433811	8	1380453043	1433811	Disconnect (2012)
	2	1454468	8	1387016442	1454468	Gravity (2013)
	2	1535109	8	1386350135	1535109	Captain Phillips (2013)Biography
	2	1675434	8	1396688981	1675434	Intouchables (2011)

Both tables already joined by matching the 'MovieId' values

Summary and References

Summary

1

Exploratory Data Analysis

- Graphical Techniques used in EDA
- Distribution in Box-Plot and Histogram

2

Statistics

- Statistics – Measures of Central Tendency
- Statistics – Measures of Dispersion

3

Outlier Treatment

- Outlier Treatments using “Clip Value” module

4

Cleaning Missing Value

- Converting data types and columns' name
- Clean the dataset without NA values
- Mathematical Operation

<https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/clip-values>

<https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/clean-missing-data>

<https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/apply-math-operation>

Summary

5

Imbalance Data

- Treat Class imbalance using SMOTE

<https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/smote>

6

Data Normalization

- Data Normalization using “Transformation Method” module

- <https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/normalize-data>

7

Joining Multiple Data

- Joining Data using “Inner Join”, “Left Outer Join”, “Left Semi-Join” or “Full Outer Join”

- <https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/join-data>

Thank you for your passion!

