Data Preparation and Exploratory Data Analysis

Citizen Analytics – An Initiative by Data Science Team

START ▶

Learning Objectives

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

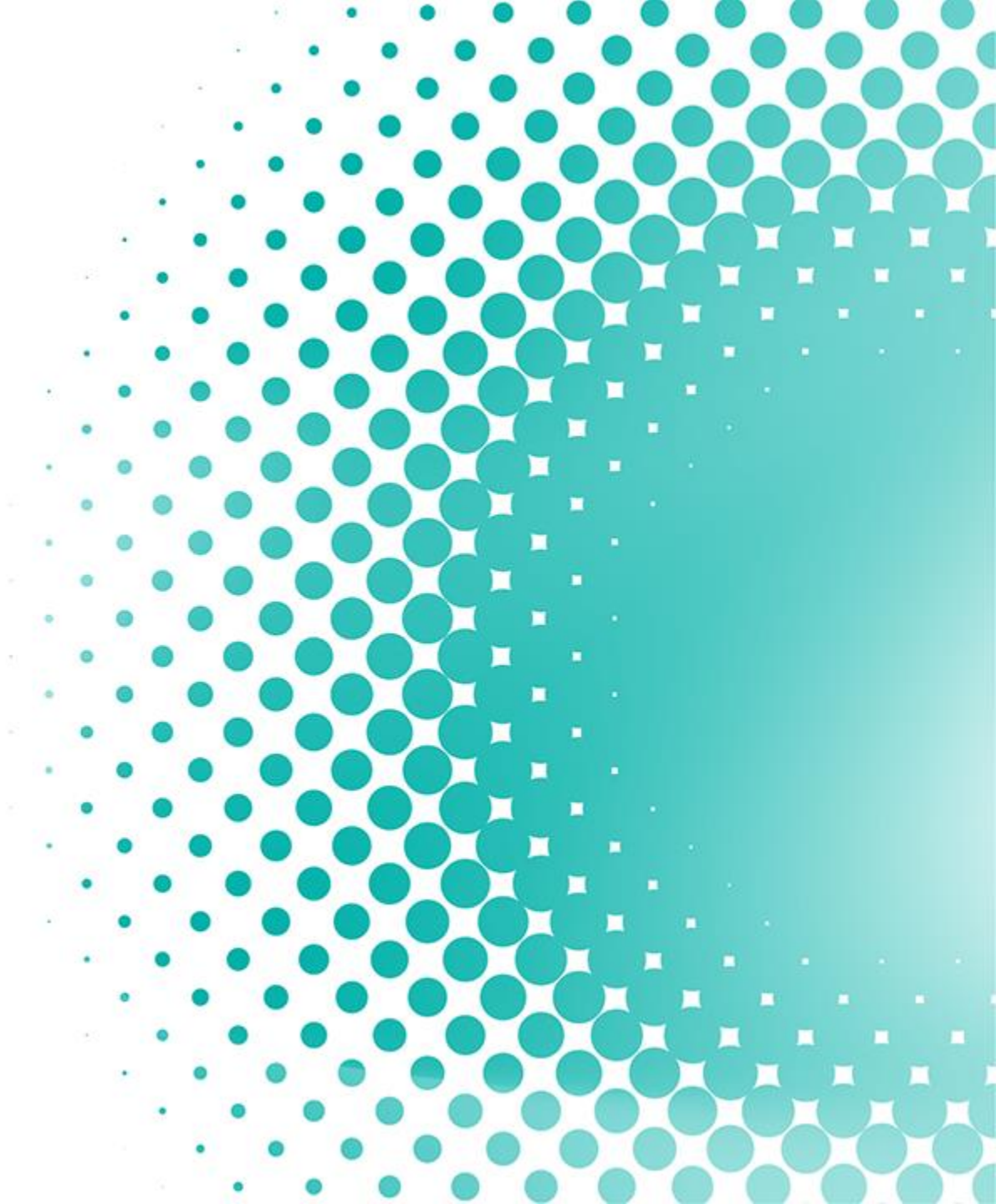
01 Understand the concept of Exploratory Data Analysis (EDA) and Statistics.

02 Identify ways to treat outliers and clean the data.

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

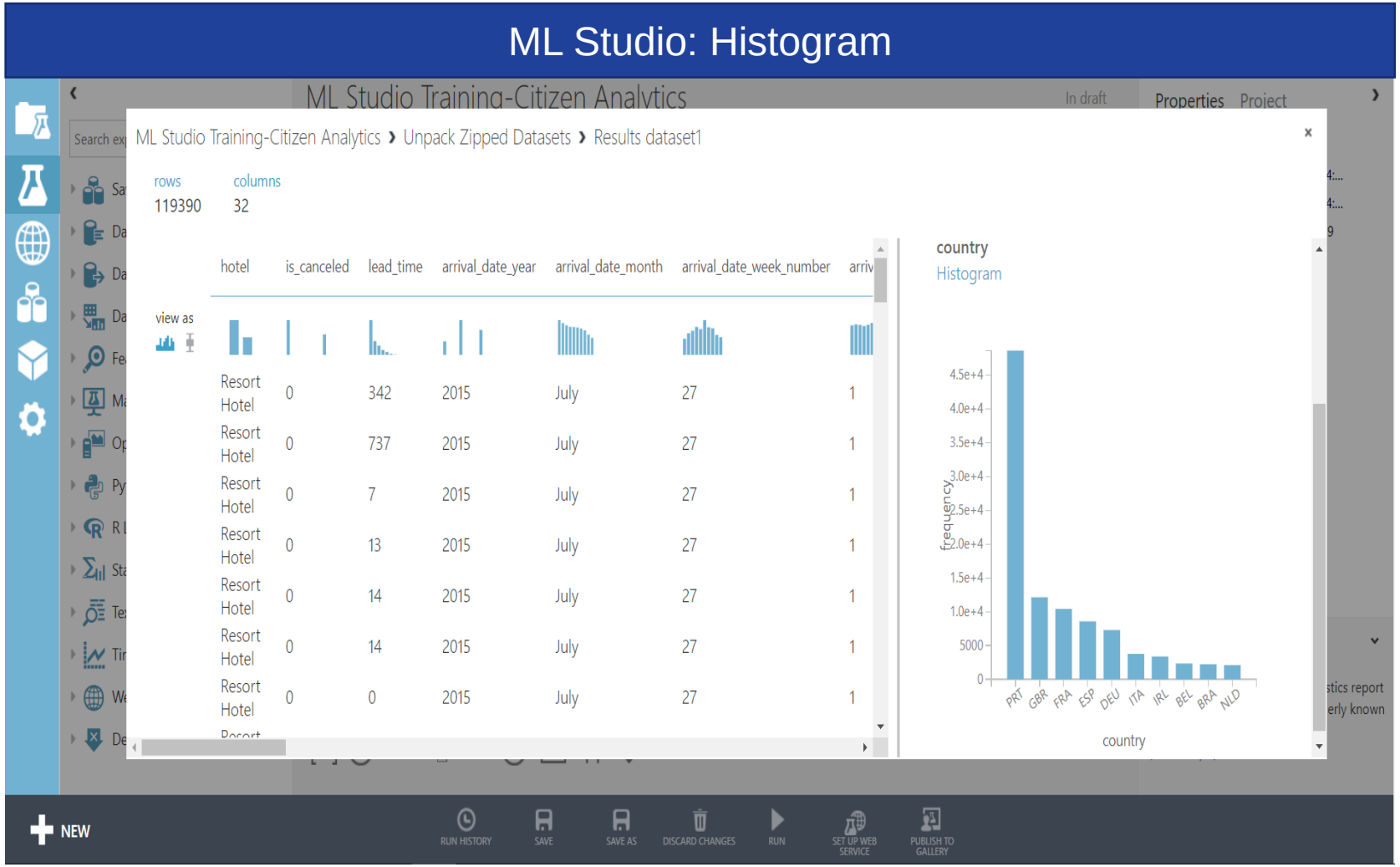
Content

Exploratory Data Analysis



PETRONAS

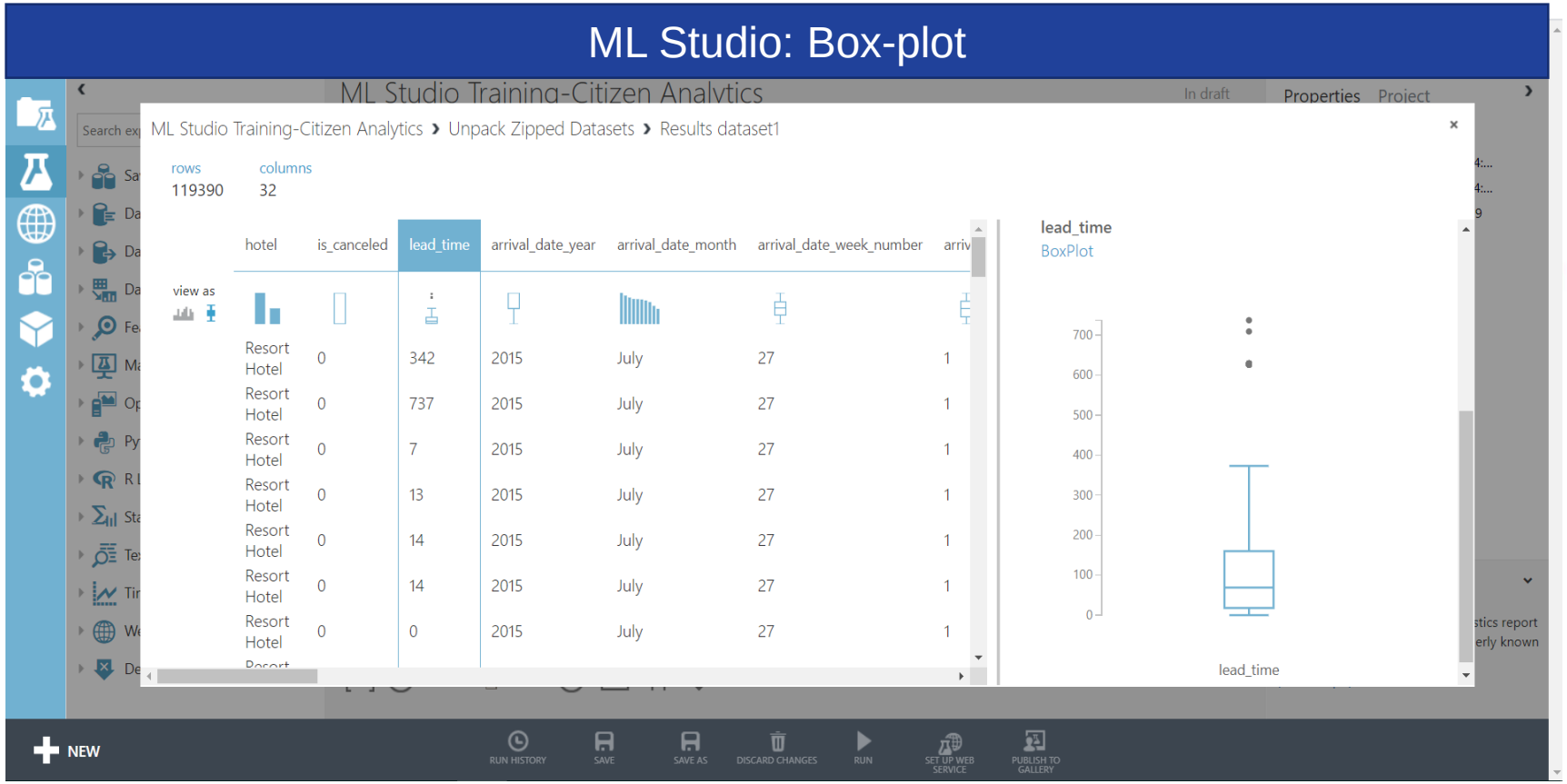
Graphical Techniques used in EDA (Histogram)

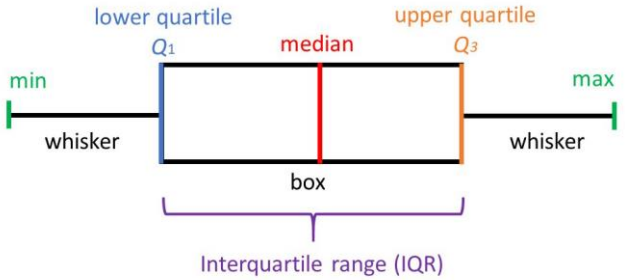


A histogram is a graphical display of data using **bars of different heights**, showing the shape and spread of continuous data. The taller the bar, the higher its occurrences.

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.

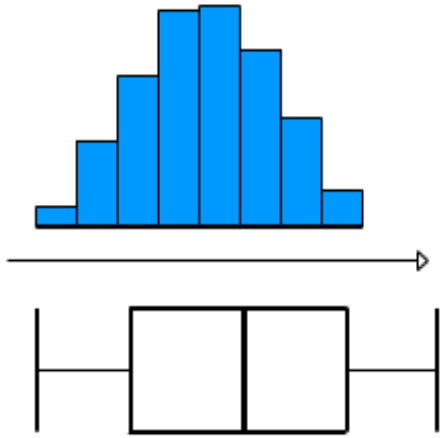




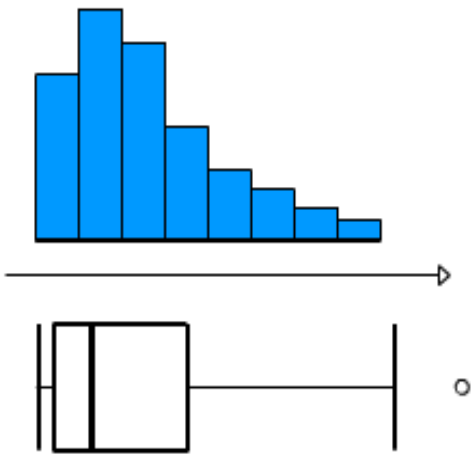
Distribution in Box-Plot and Histogram

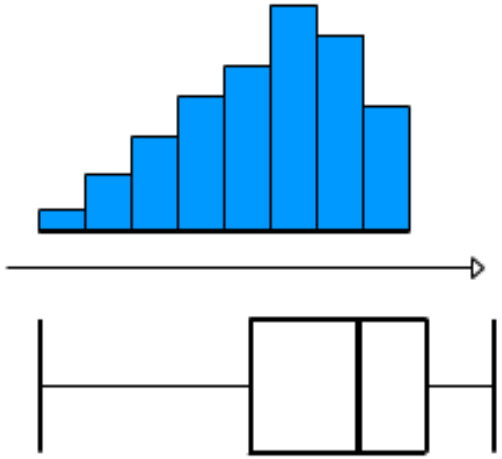
- In symmetric distributions (normal), the mean, median, and mode are the same.
- In skewed data, the mean and median lie further toward the skew than the mode, either skewed to the right or left.

Normally distributed



Skewed





Statistics

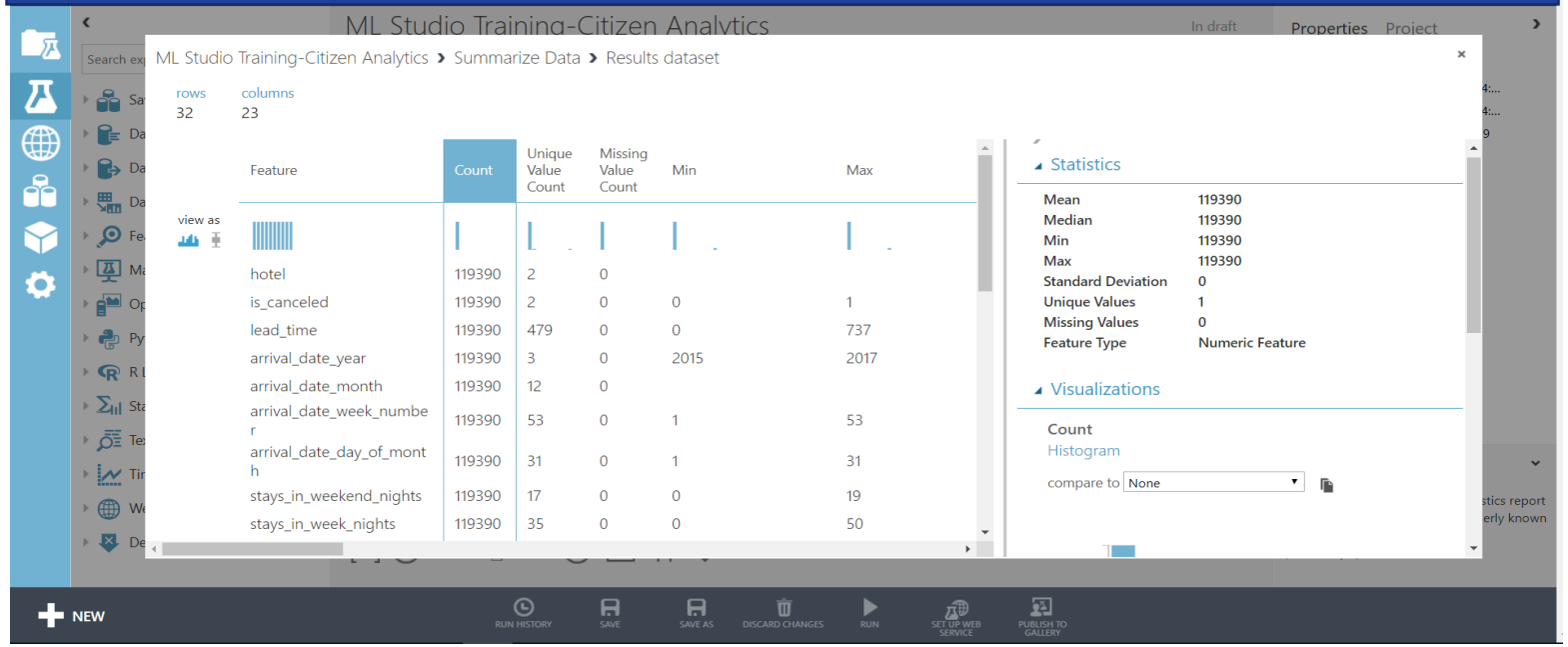
PETRONAS

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

ML Studio : Mean, Median and 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

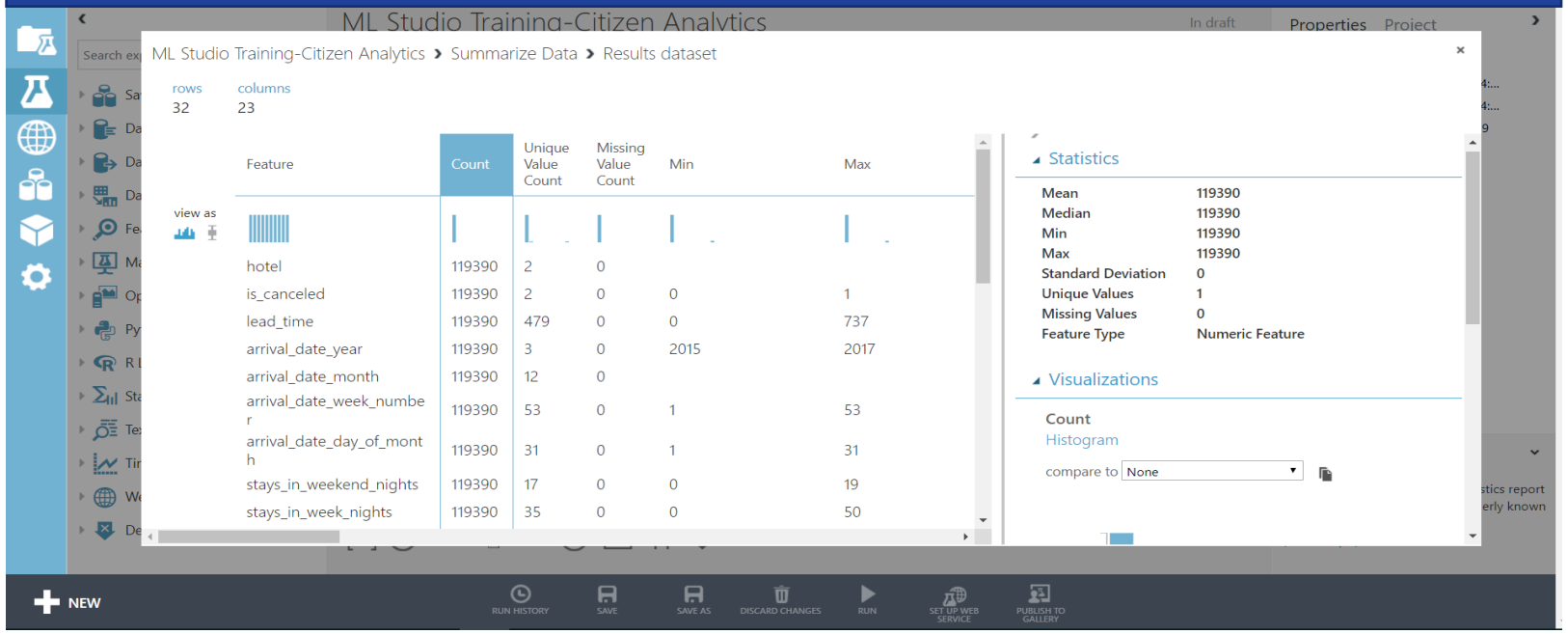
PETRONAS

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

ML Studio : Variance and Standard Deviation



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

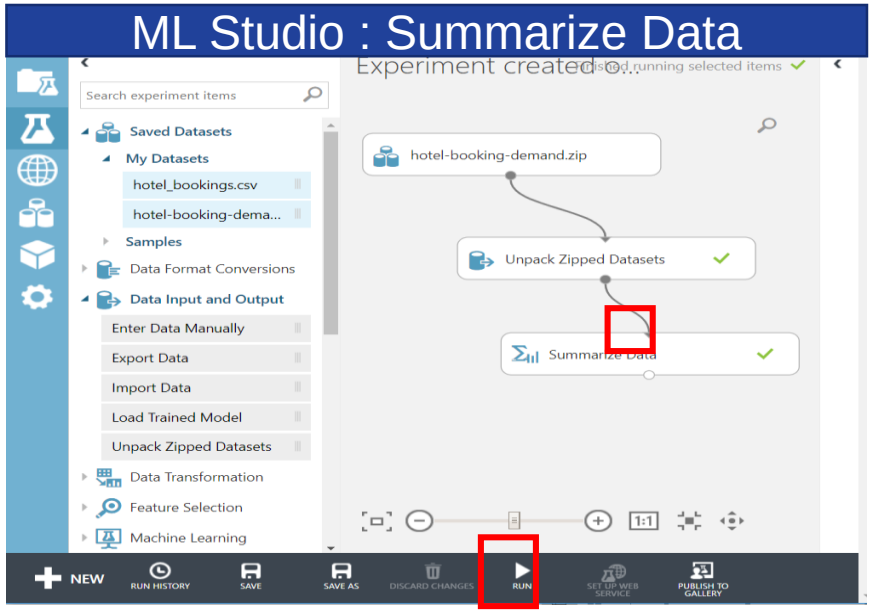
Standard deviation is just a square root of the variance, measuring how disperse is the data from the mean.

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

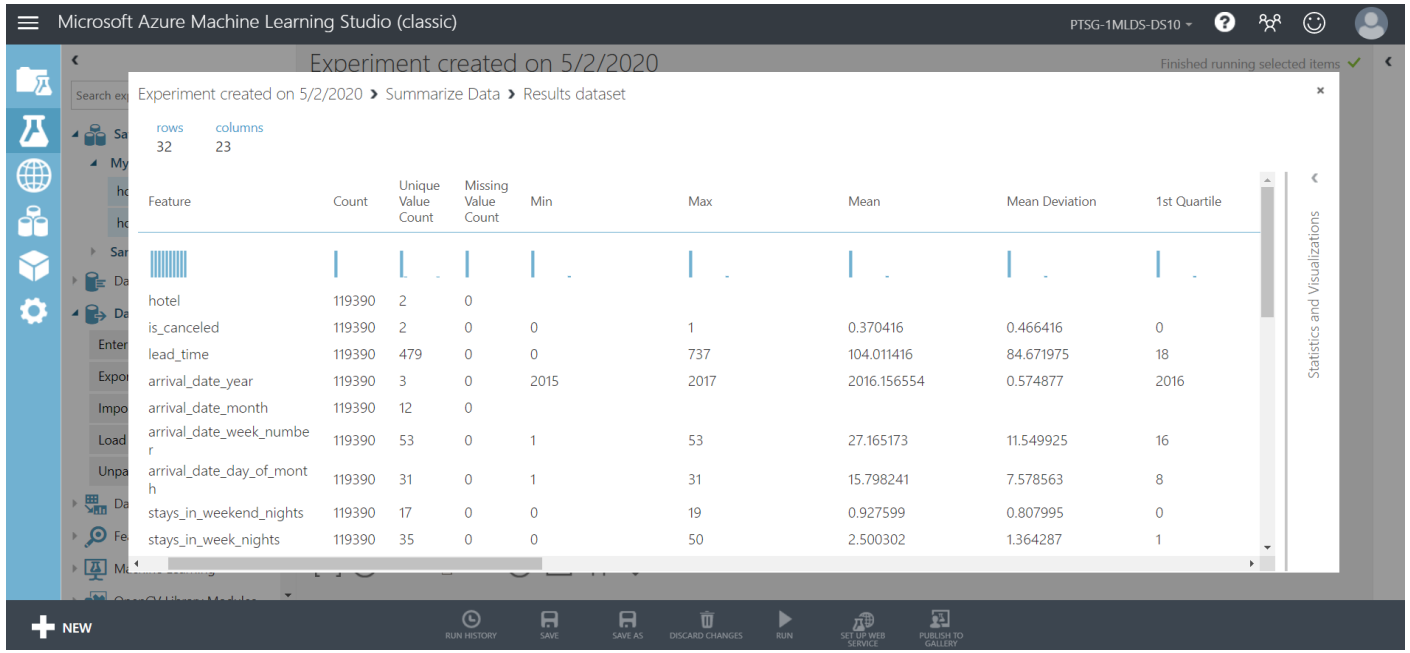
PETRONAS

Summary of the data

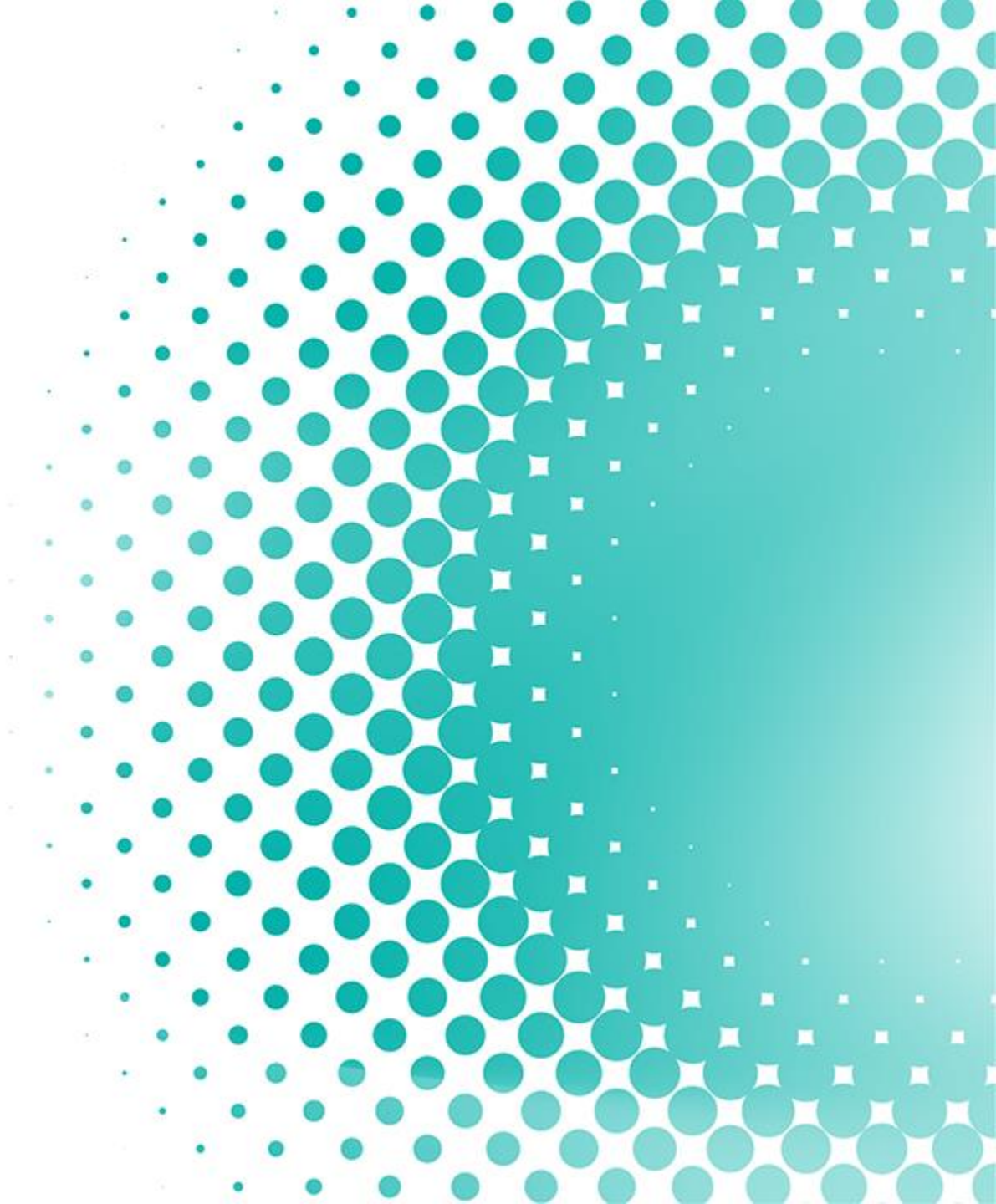
‘Summarize Data’ module allows us to see all related statistical measures such as mean, median, quartiles etc in just one pop out window.



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

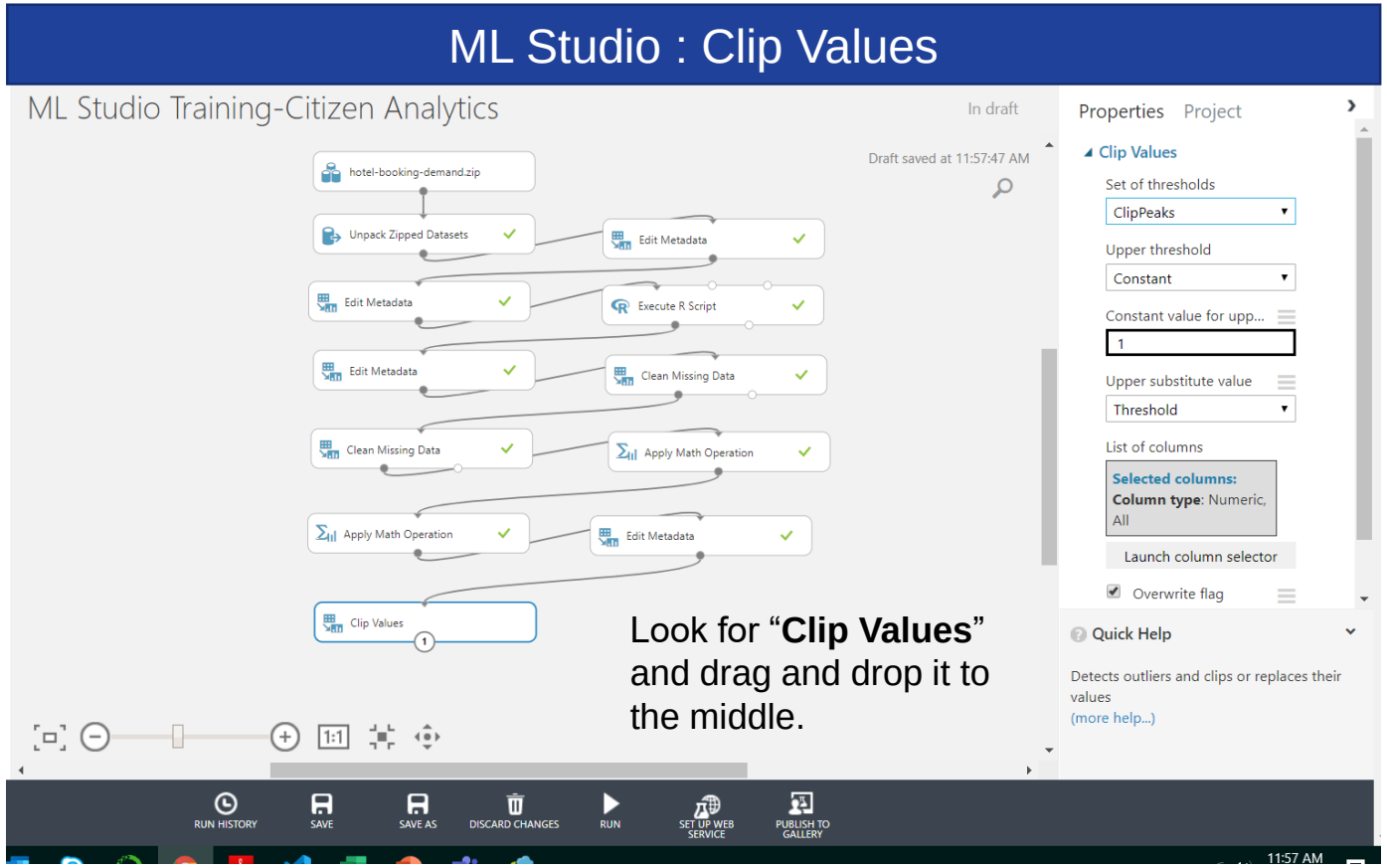


Outlier Treatment



Outliers treatment

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



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

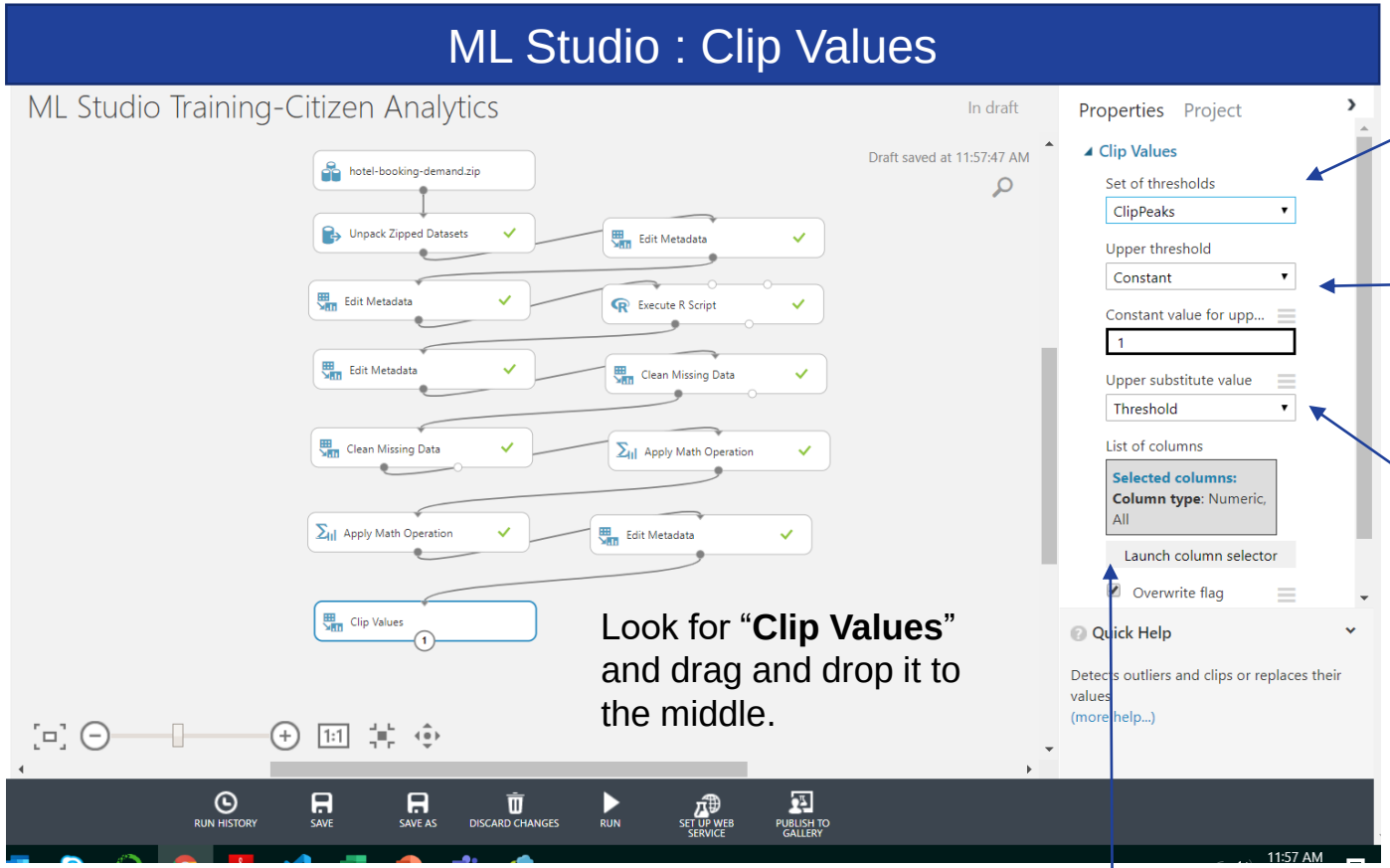
1 "**ClipPeaks**" is clipping the higher threshold. Choose this if your variable has more outliers at the upper level of the value of the variable

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



"**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.

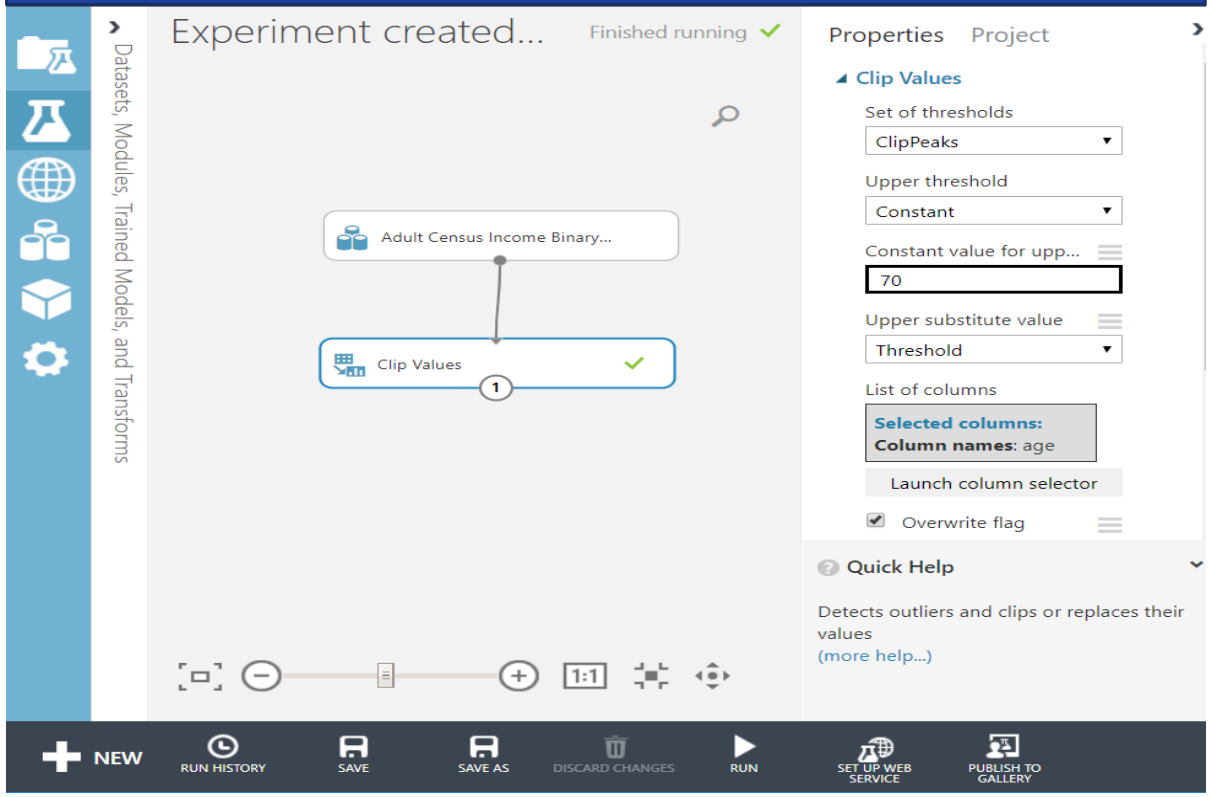
Lastly, select which column you want to treat for its outliers.

Outliers treatment - Example

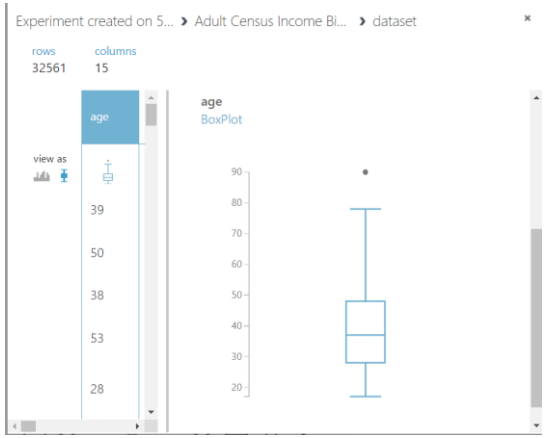
Dataset = Adult Census Income Binary

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

ML Studio : Clip Values



Before



The dot represents the outliers in this data

After

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

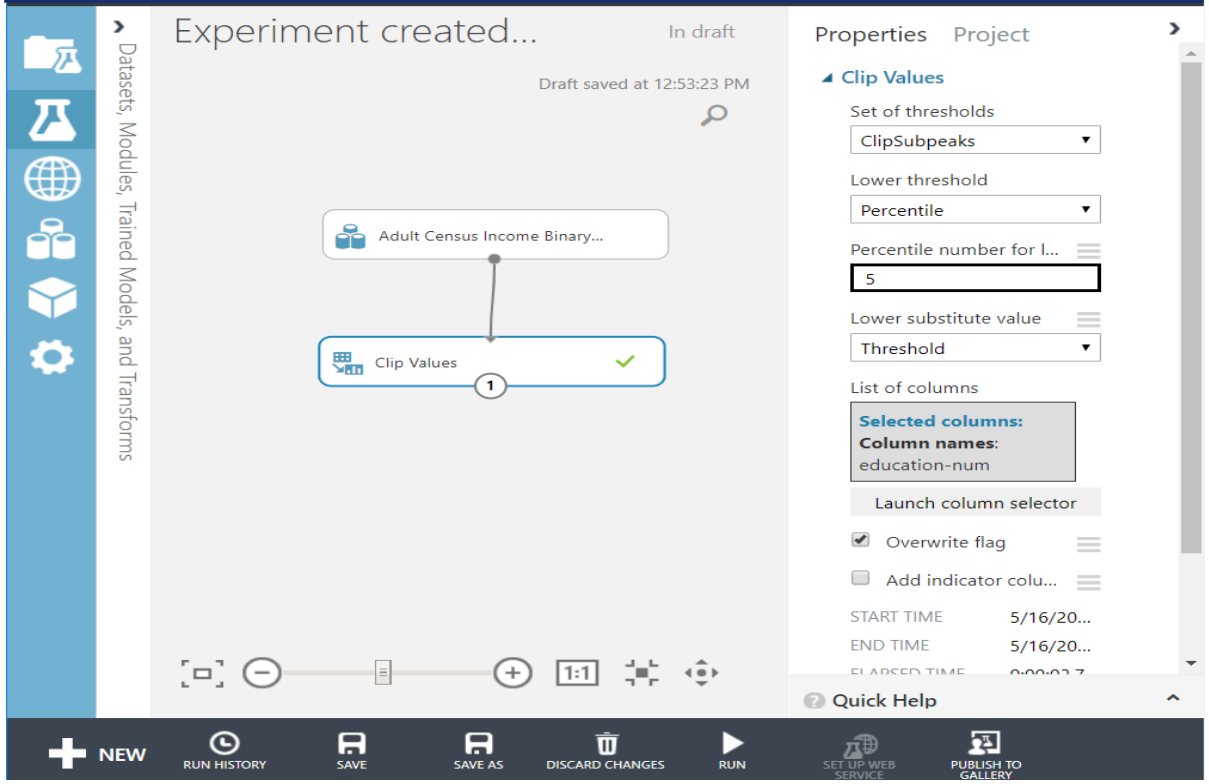
PETRONAS

Outliers treatment - Example

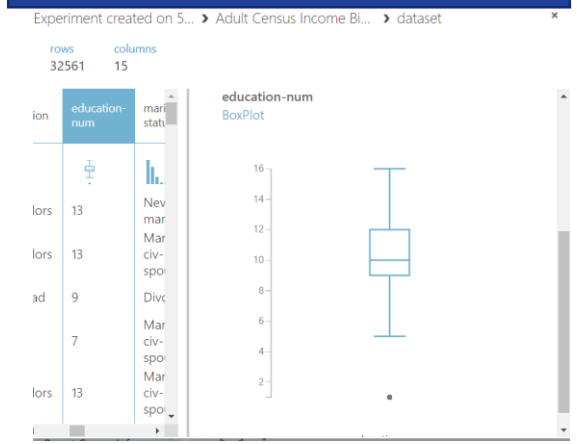
Dataset = Adult Census Income Binary

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

ML Studio : Clip Values

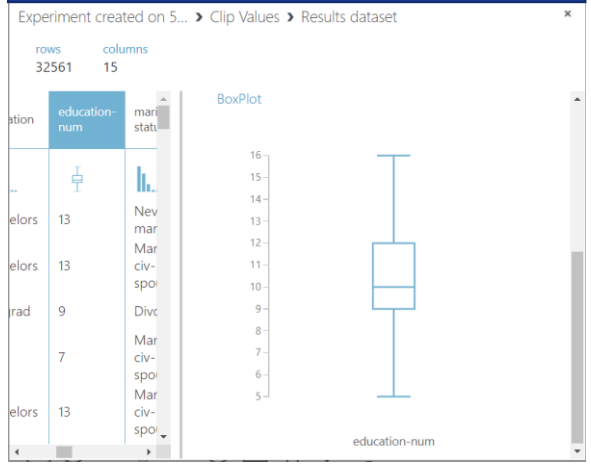


Before



The dot represents the outliers in this data

After

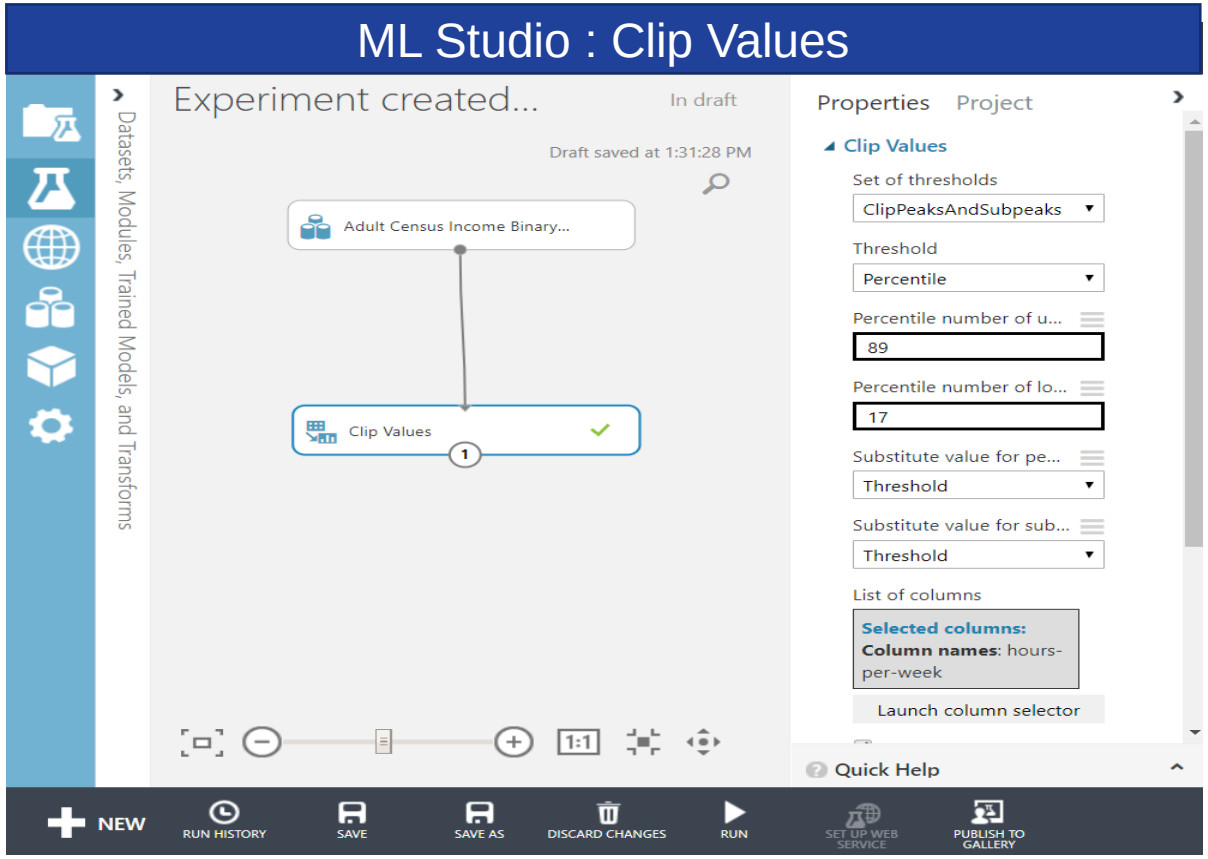


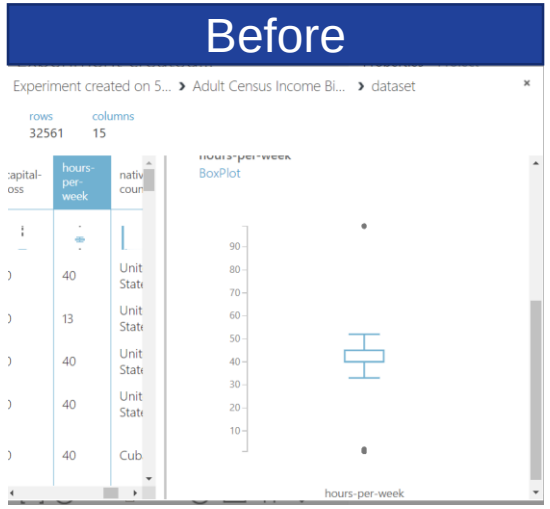
The data is already clipped and the outliers is no longer there as the value is replaced by using the new lower threshold.

Outliers treatment - Example

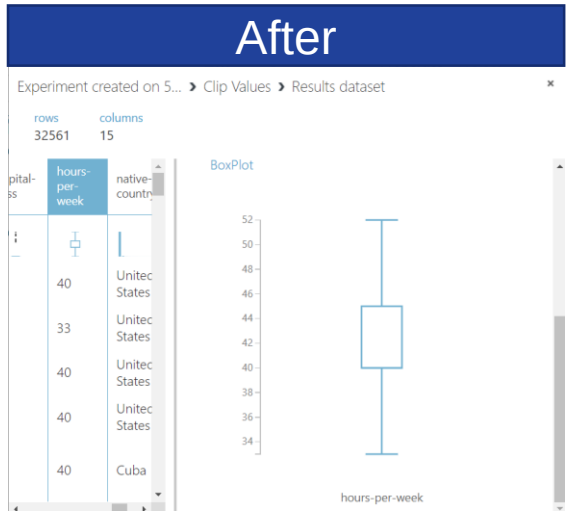
Dataset = Adult Census Income Binary

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)





The dot represents the outliers in this data



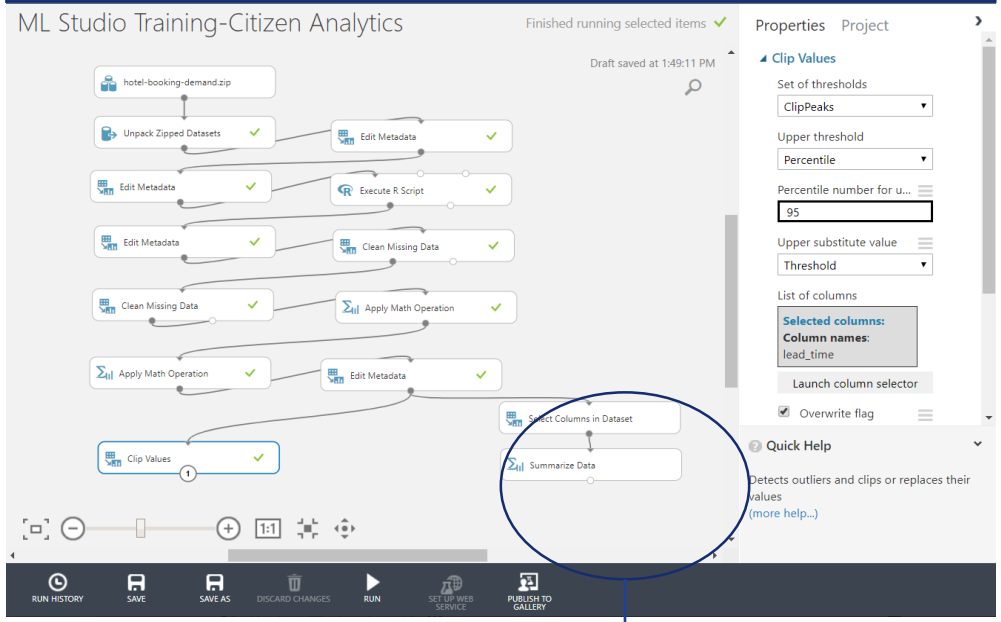
The data is already clipped and the outliers is no longer there as the value is replaced by using the new upper and lower percentile of threshold.

PETRONAS

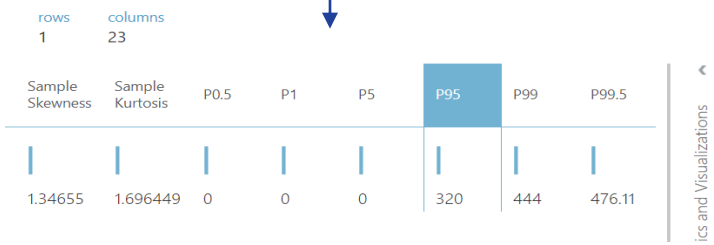
Outliers treatment

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

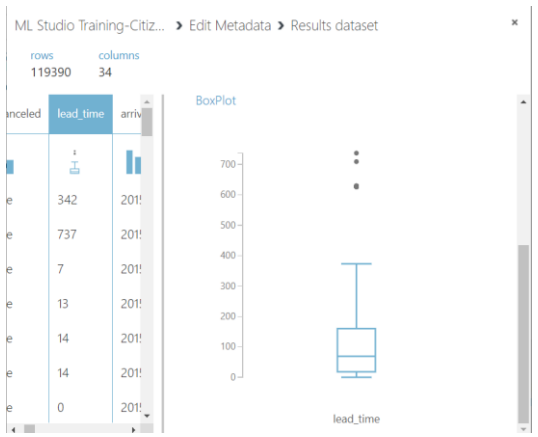
ML Studio : Clip Values



Result

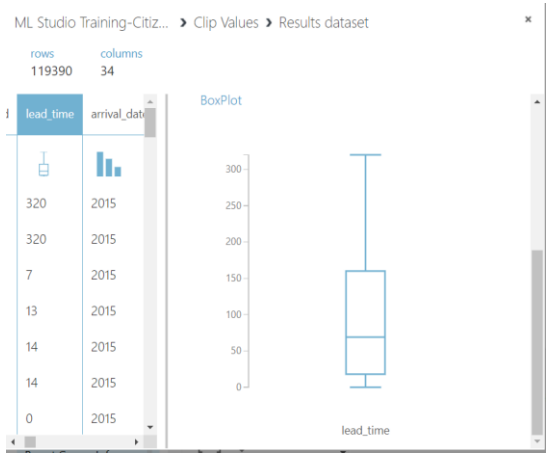


Before



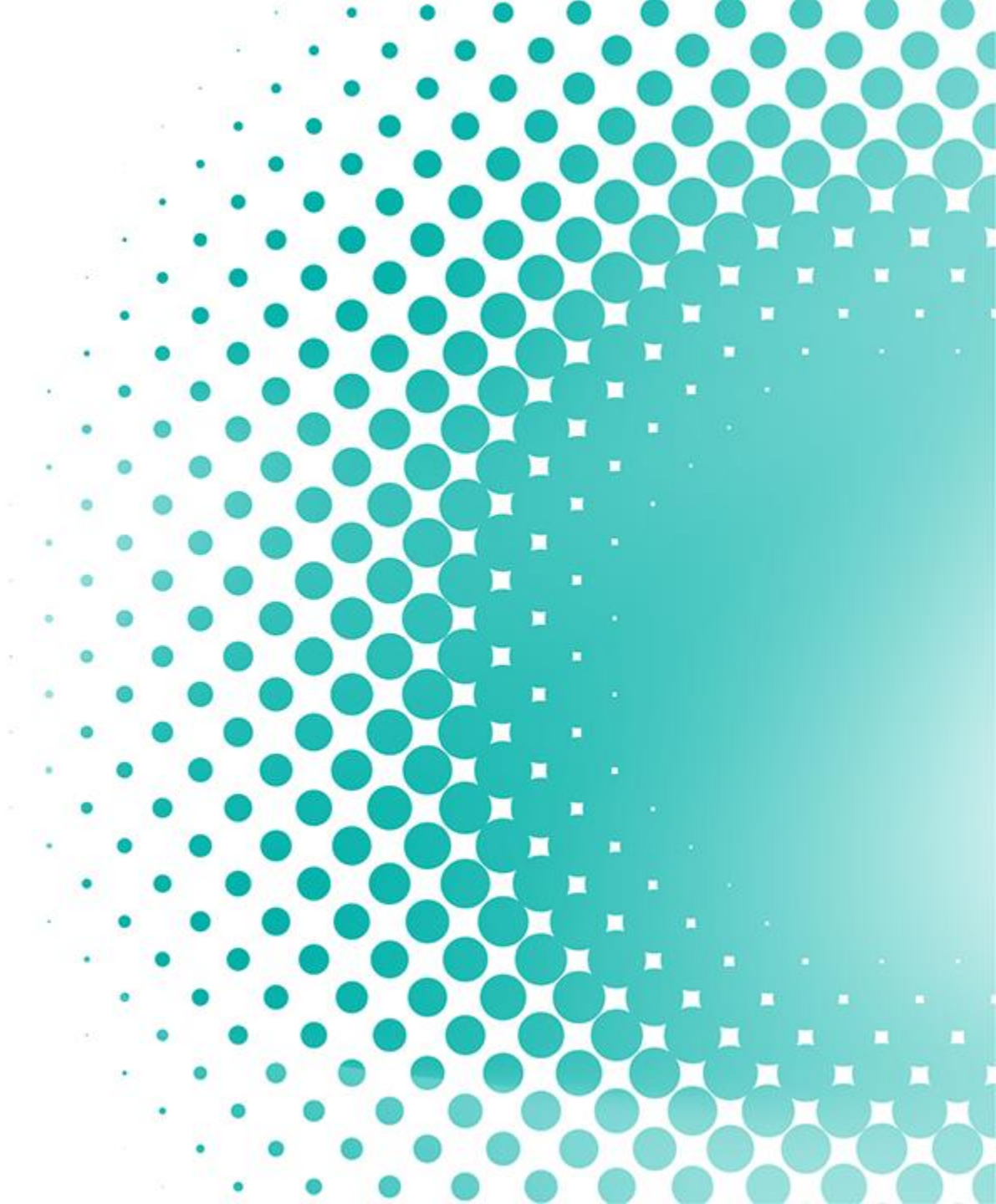
The dot represents the outliers in this data

After



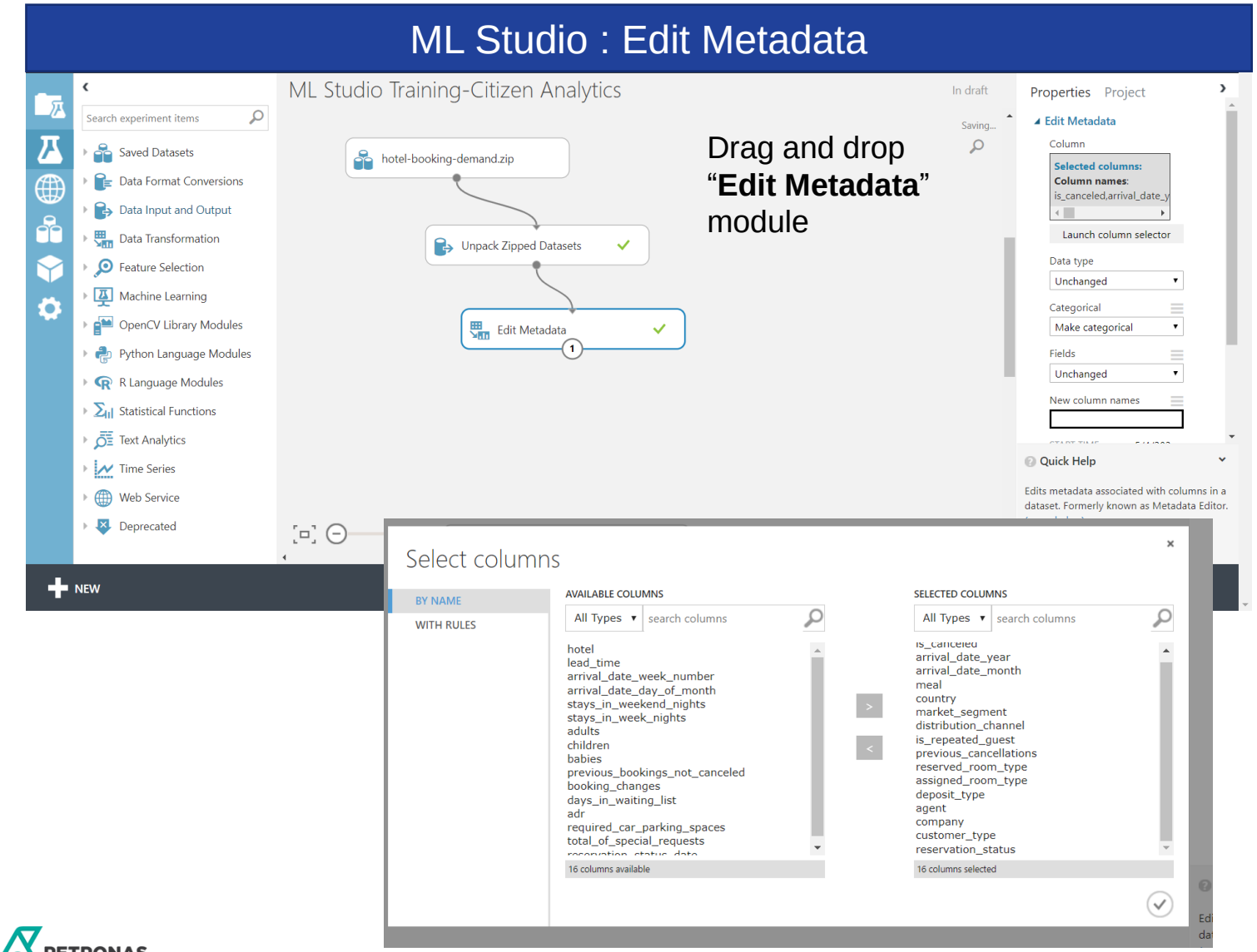
The data is already clipped and the outliers is no longer there as the value is replaced by using the new upper threshold.

Cleaning Missing Value



PETRONAS

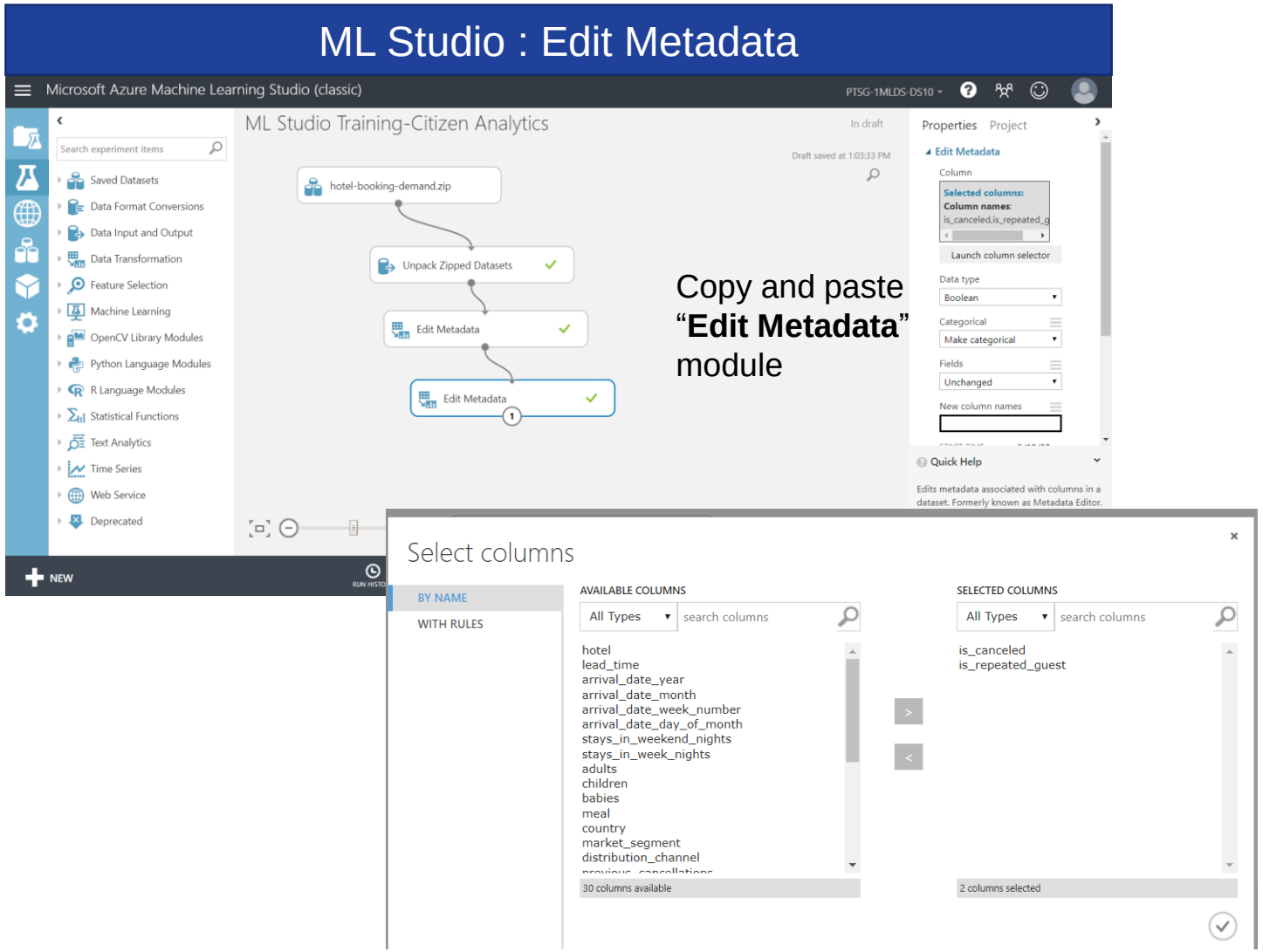
Converting data types and columns' name



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.

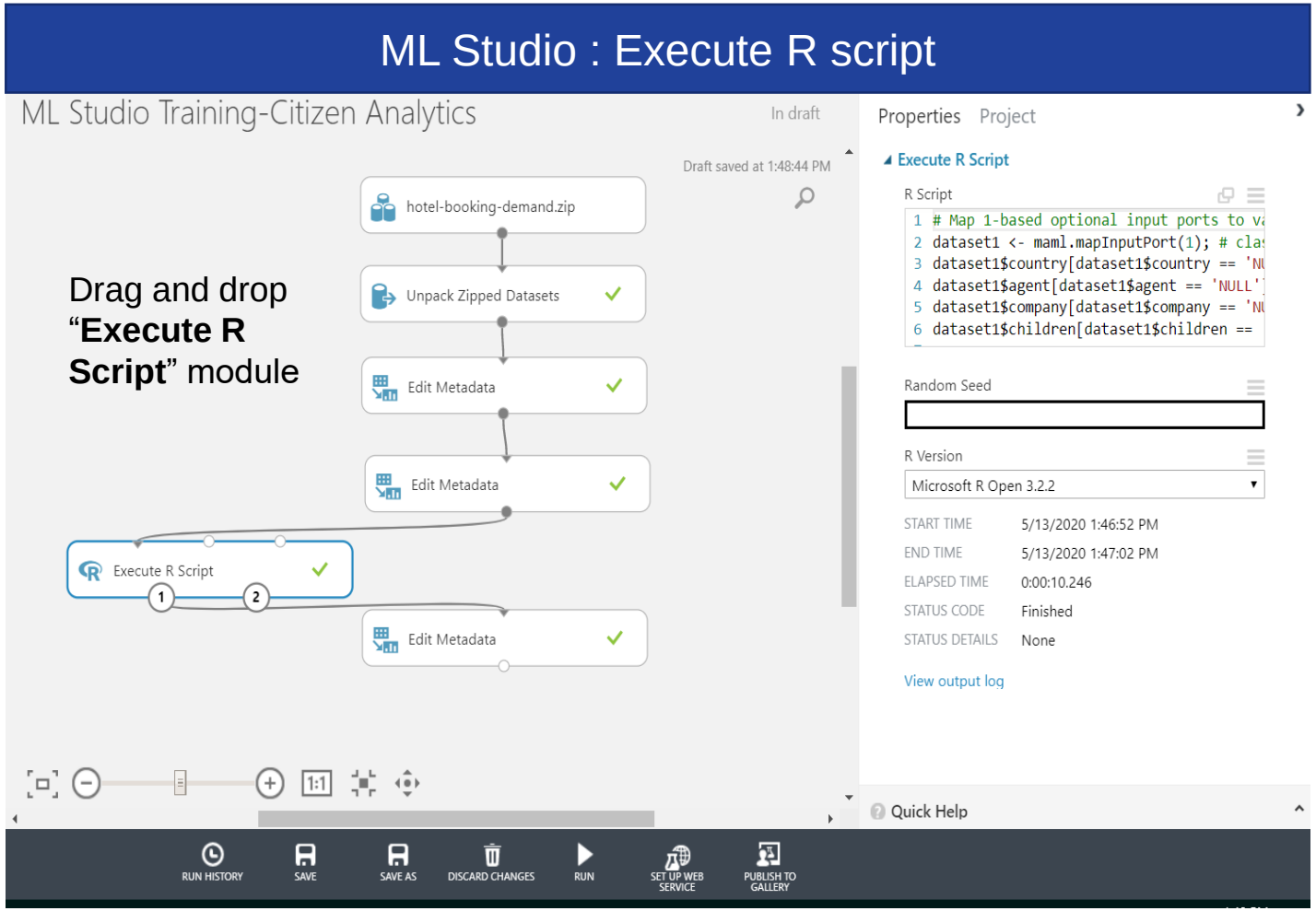
Converting data types and columns' name



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 and 'Boolean' for datatype.

Clean the dataset without NA values

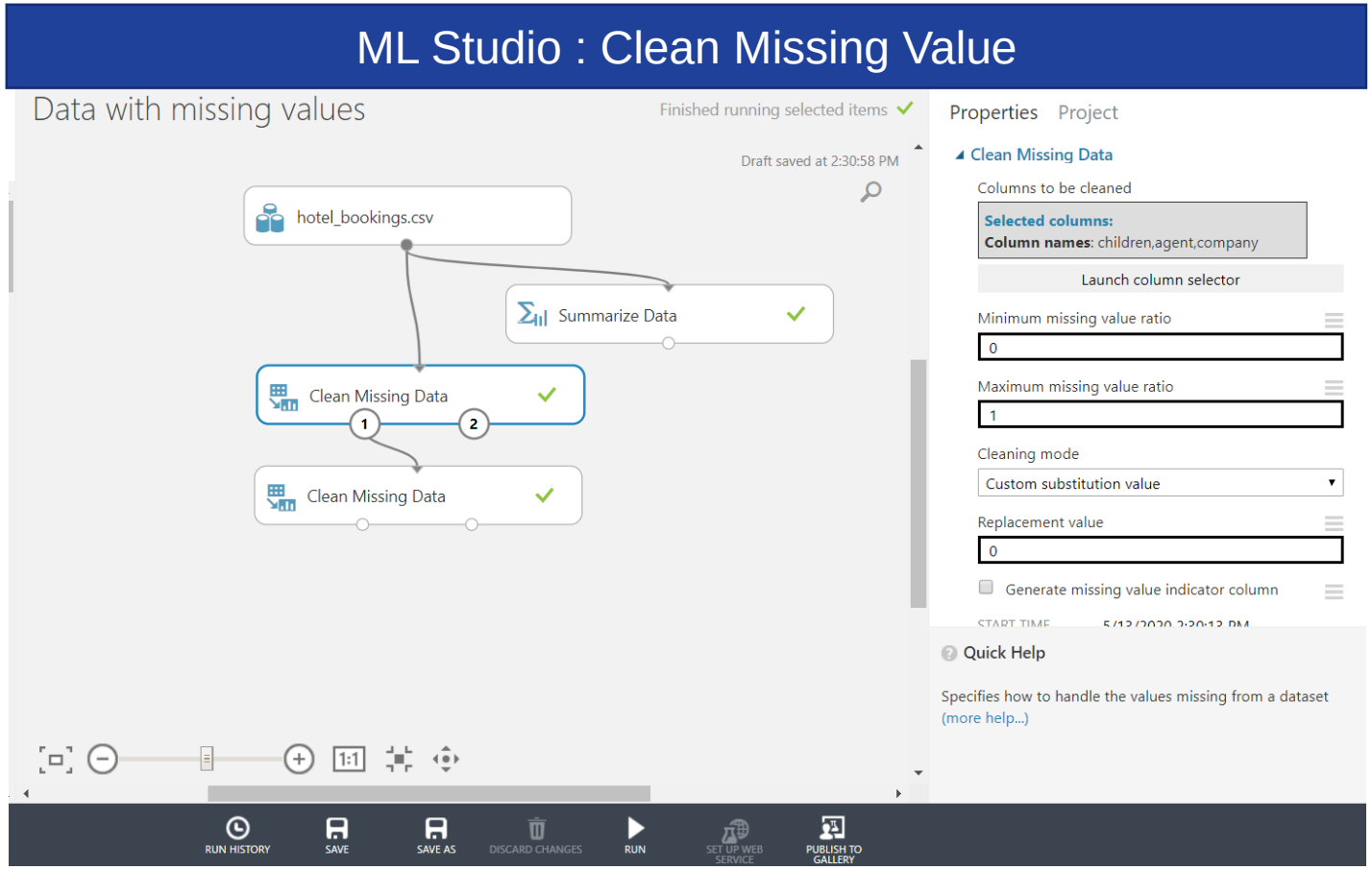


Change "**NULL**" and "**NA**" values in 'Children', 'Agent', 'Company' and 'Country' to NA

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

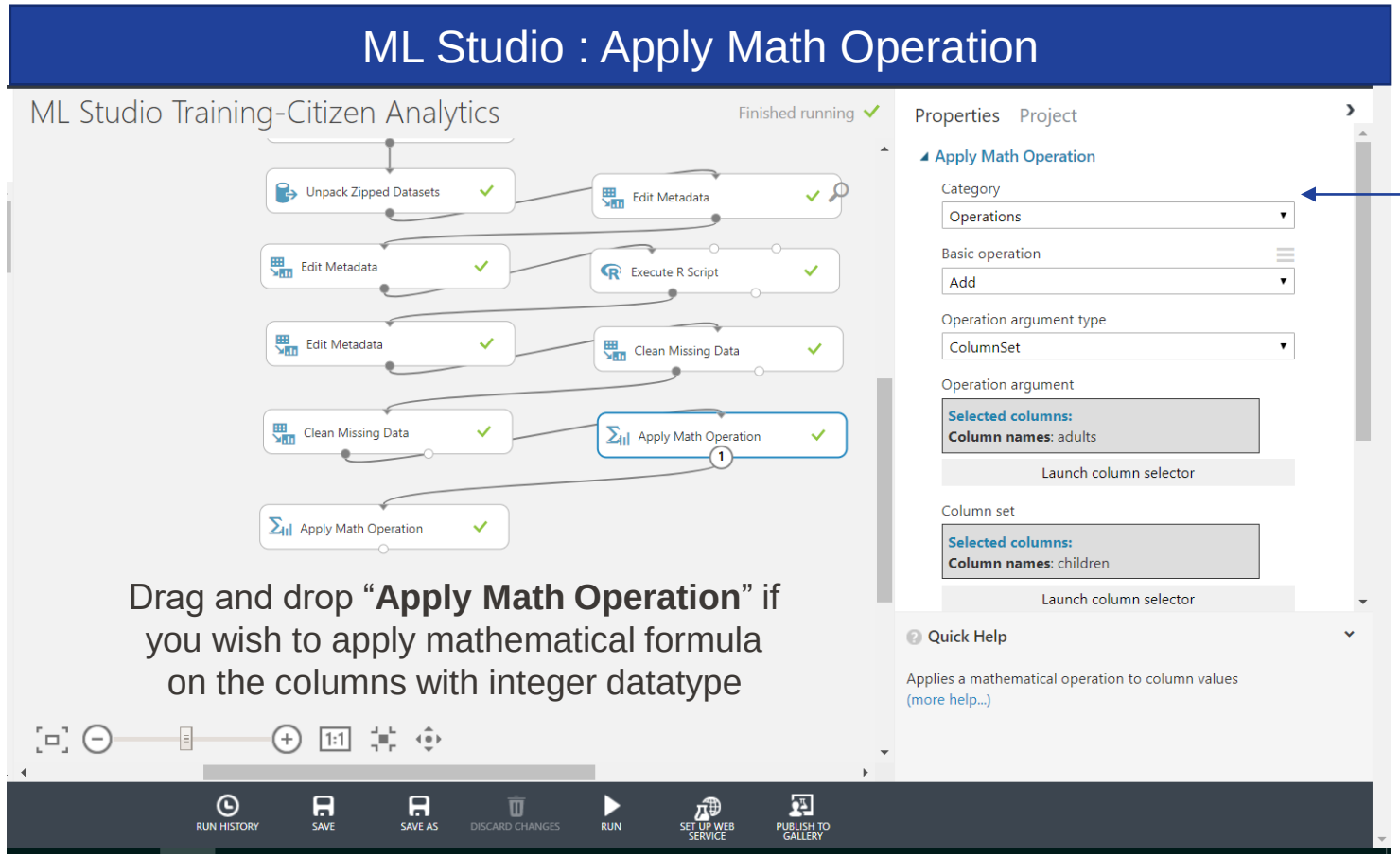
```
# Map 1-based optional input ports to variables
dataset1 <- maml.mapInputPort(1); # class: data.frame
dataset1$country[dataset1$country == 'NULL'] = NA;
dataset1$agent[dataset1$agent == 'NULL'] = NA;
dataset1$company[dataset1$company == 'NULL'] = NA;
dataset1$children[dataset1$children == 'NA'] =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

Clean the dataset with NA values



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

Mathematical Operation



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

PETRONAS

Class Imbalance

Imbalanced data happen when the classes are not represented equally

ML Studio : Imbalance Data

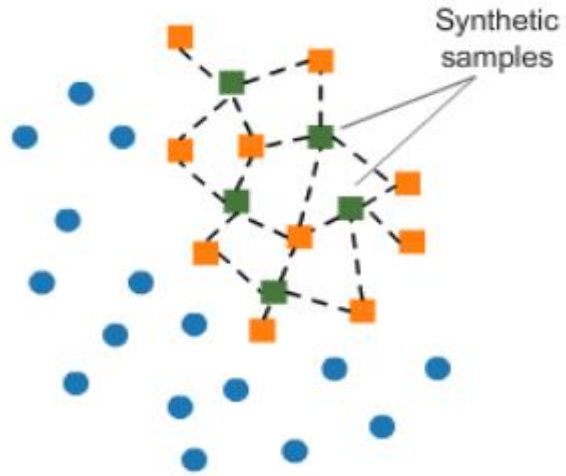
ML Studio Training-Citiz... › Clip Values › Results dataset

rows	columns
119390	34

el	is_repeated_guest	p
	False	0
	False	0
	False	0
	False	0
	False	0
	False	0
	False	0
	False	0

Dealing with imbalanced dataset using SMOTE – Synthetic Minority Oversampling Technique

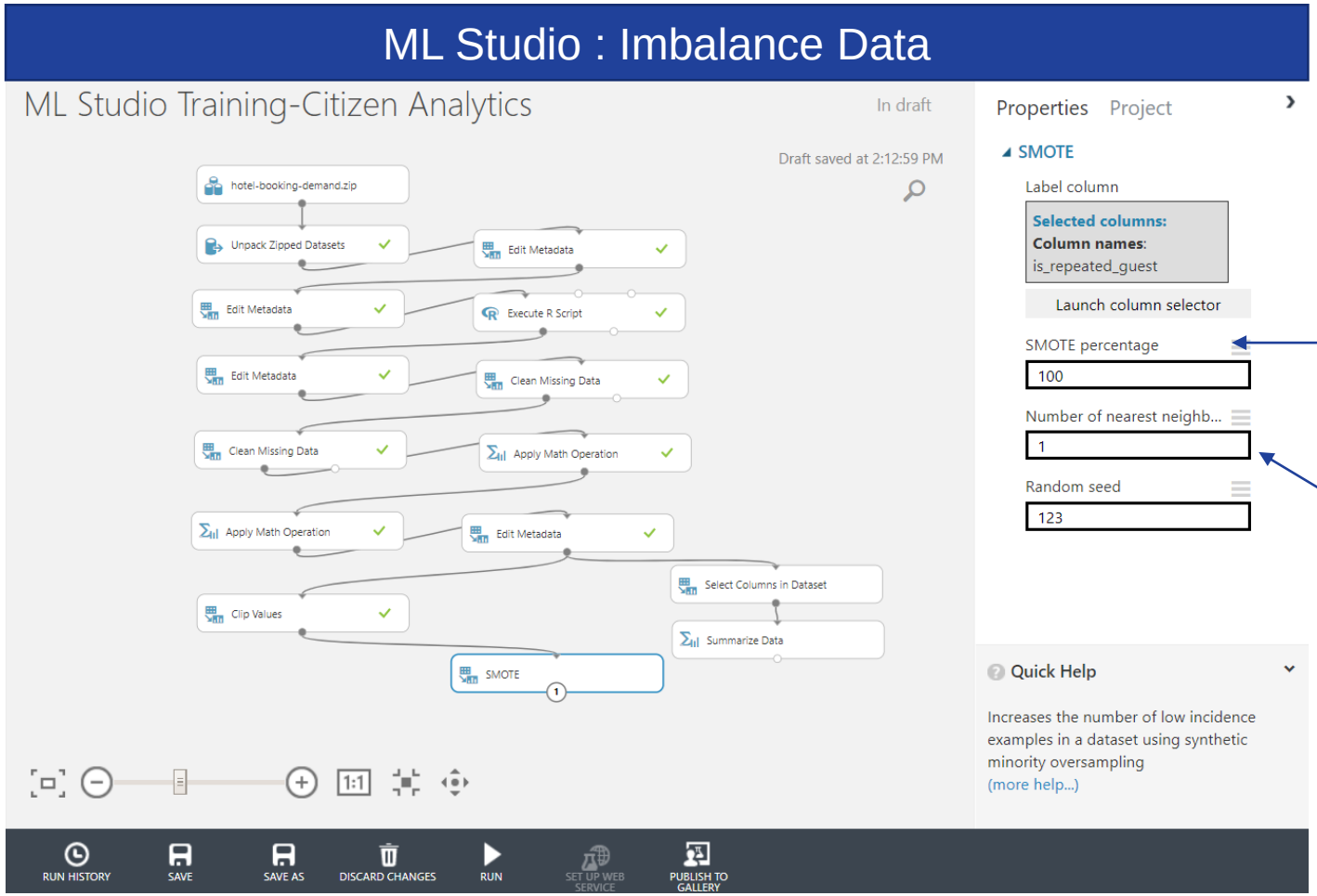
SMOTE is just creating new synthetics datapoints between two datapoints



PETRONAS

Class Imbalance

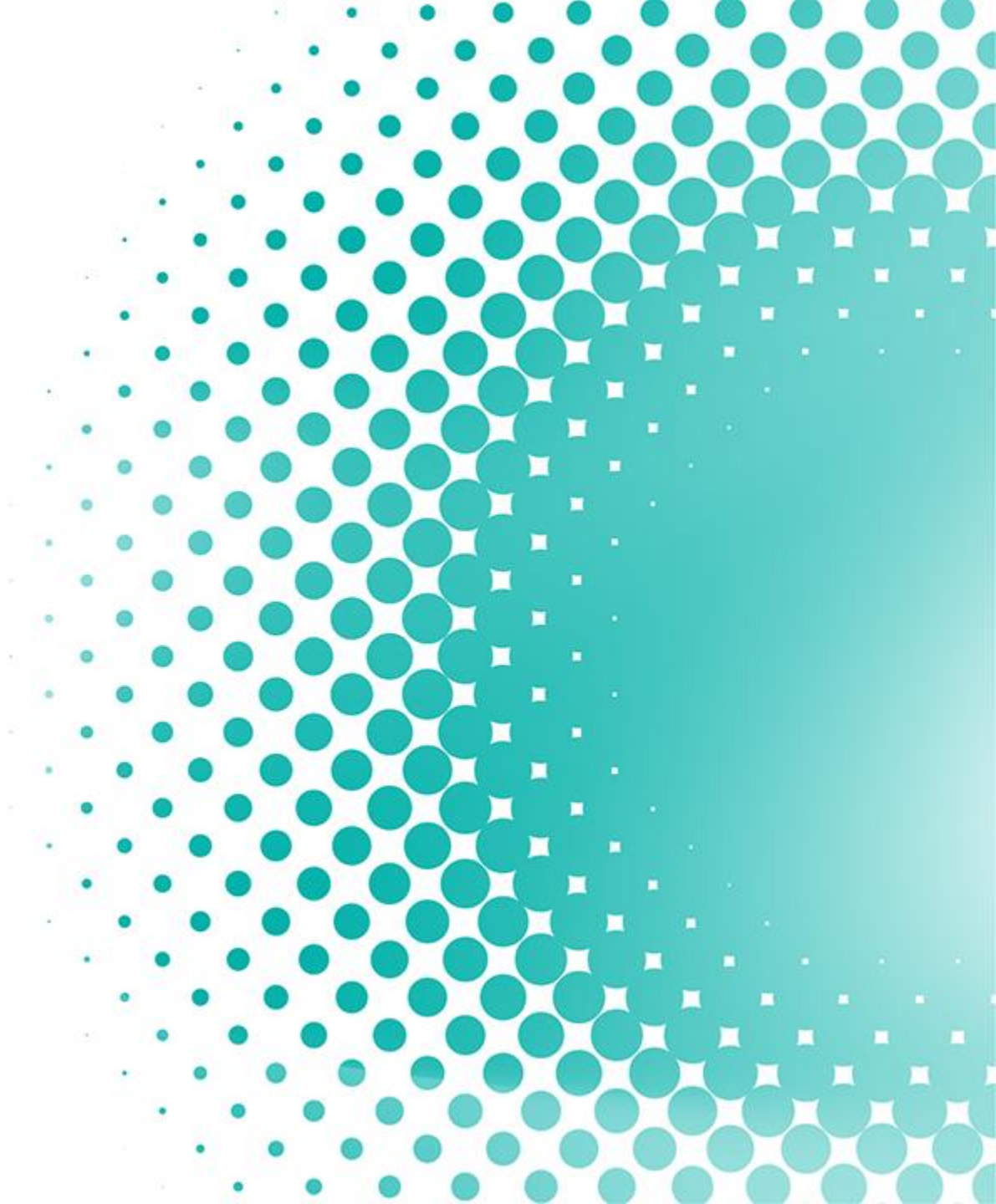
Dealing with imbalance data by oversampling method



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



PETRONAS

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

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

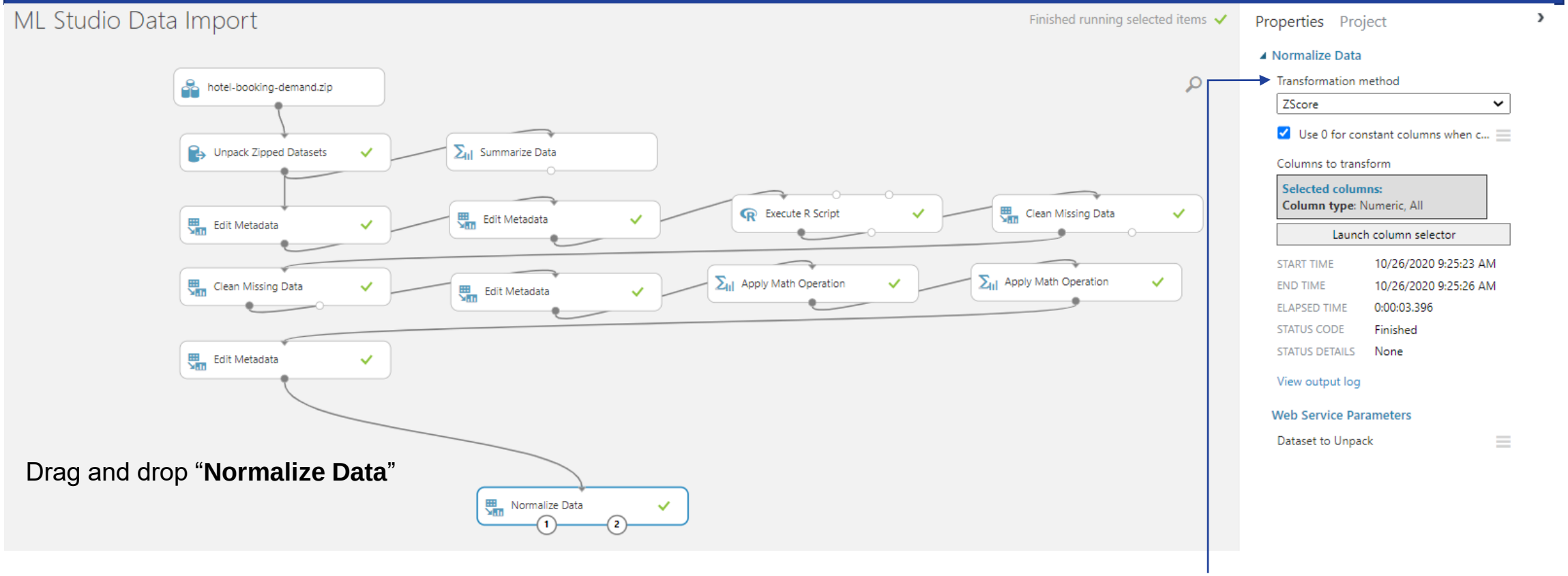
LogNormal Normalization

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

TanH Normalization

Data Normalization - Example

ML Studio : Normalize Data



Drag and drop "**Normalize Data**"

At "**Transformation method**", select type of normalization and select columns

Data Normalization - Example

Before

ML Studio Data Import › Edit Metadata › Results dataset

rows 119390 columns 34

hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_d
Resort Hotel	False	342	2015	July	27	1
Resort Hotel	False	737	2015	July	27	1
Resort Hotel	False	7	2015	July	27	1
Resort Hotel	False	13	2015	July	27	1
Resort Hotel	False	14	2015	July	27	1
Resort Hotel	False	14	2015	July	27	1
Resort Hotel	False	0	2015	July	27	1
Resort Hotel	False	9	2015	July	27	1
Resort Hotel	True	85	2015	July	27	1
Resort Hotel	True	75	2015	July	27	1
Resort Hotel	True	23	2015	July	27	1
Resort Hotel	False	35	2015	July	27	1
Resort Hotel	False	68	2015	July	27	1

The original data with values ranges in hundred.

After

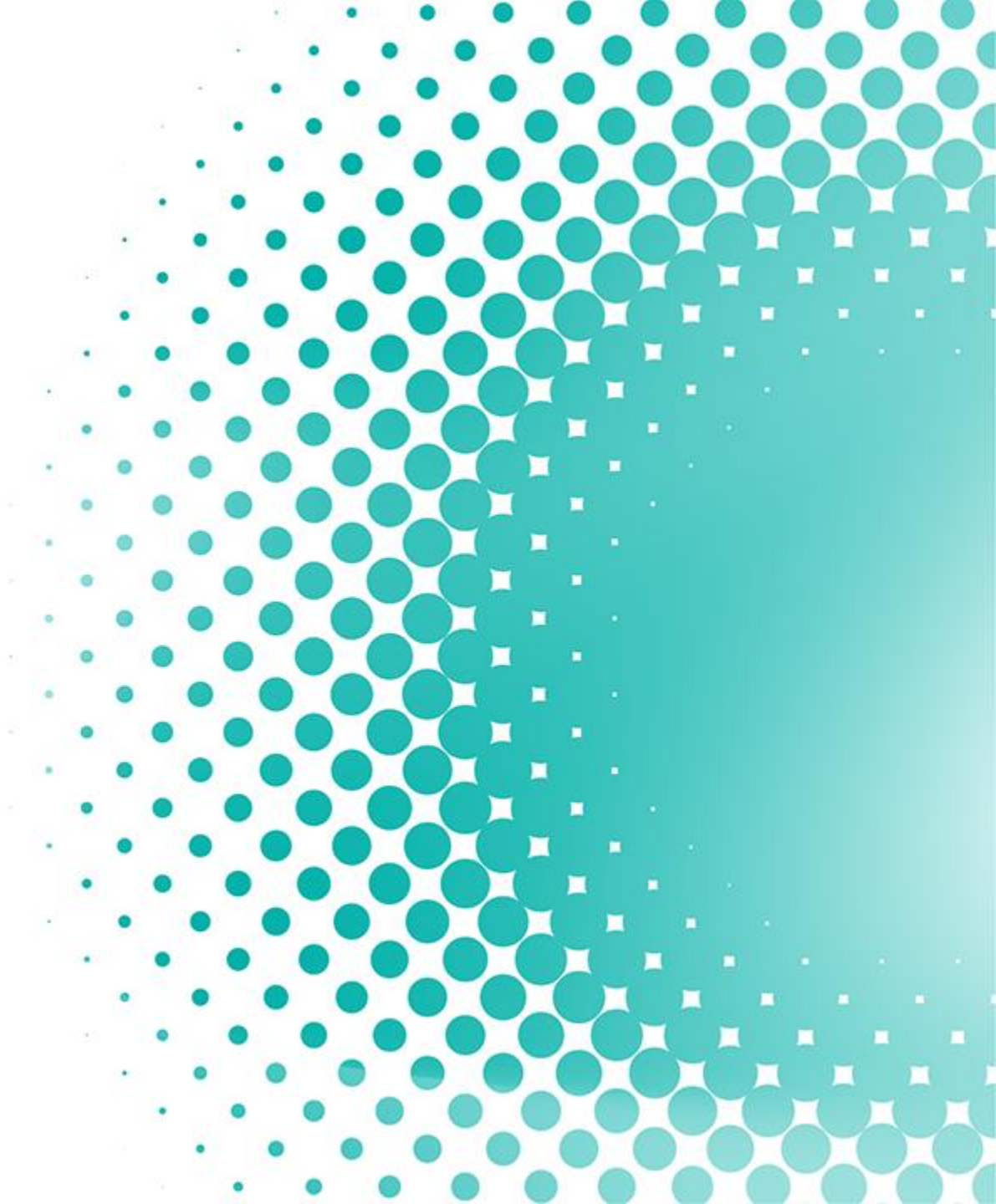
ML Studio Data Import › Normalize Data › Transformed dataset

rows 119390 columns 34

hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_
Resort Hotel	False	2.227051	2015	July	-0.012141	-1.685297
Resort Hotel	False	5.923385	2015	July	-0.012141	-1.685297
Resort Hotel	False	-0.907814	2015	July	-0.012141	-1.685297
Resort Hotel	False	-0.851667	2015	July	-0.012141	-1.685297
Resort Hotel	False	-0.842309	2015	July	-0.012141	-1.685297
Resort Hotel	False	-0.842309	2015	July	-0.012141	-1.685297
Resort Hotel	False	-0.973319	2015	July	-0.012141	-1.685297
Resort Hotel	False	-0.889098	2015	July	-0.012141	-1.685297
Resort Hotel	True	-0.177905	2015	July	-0.012141	-1.685297
Resort Hotel	True	-0.271483	2015	July	-0.012141	-1.685297
Resort Hotel	True	-0.758089	2015	July	-0.012141	-1.685297
Resort Hotel	False	-0.645795	2015	July	-0.012141	-1.685297
Resort Hotel	False	-0.336988	2015	July	-0.012141	-1.685297

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

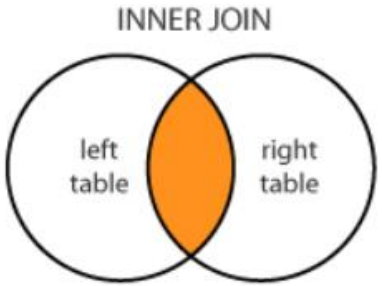
Joining Multiple Data

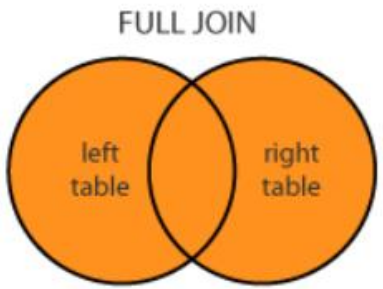


PETRONAS

Joining data

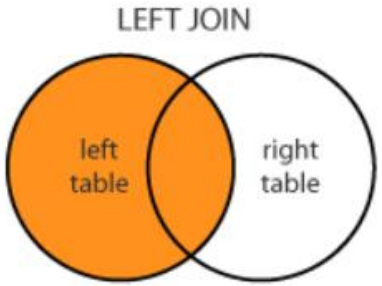
Data is merged on the common column on both tables

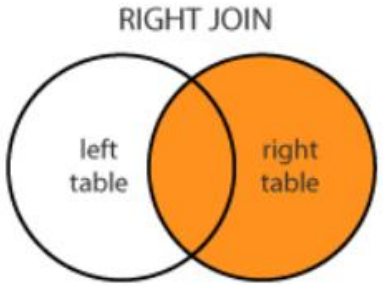




Data is combined regardless of the columns name.

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

view as

UserId	MovieId	Rating	Timestamp
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

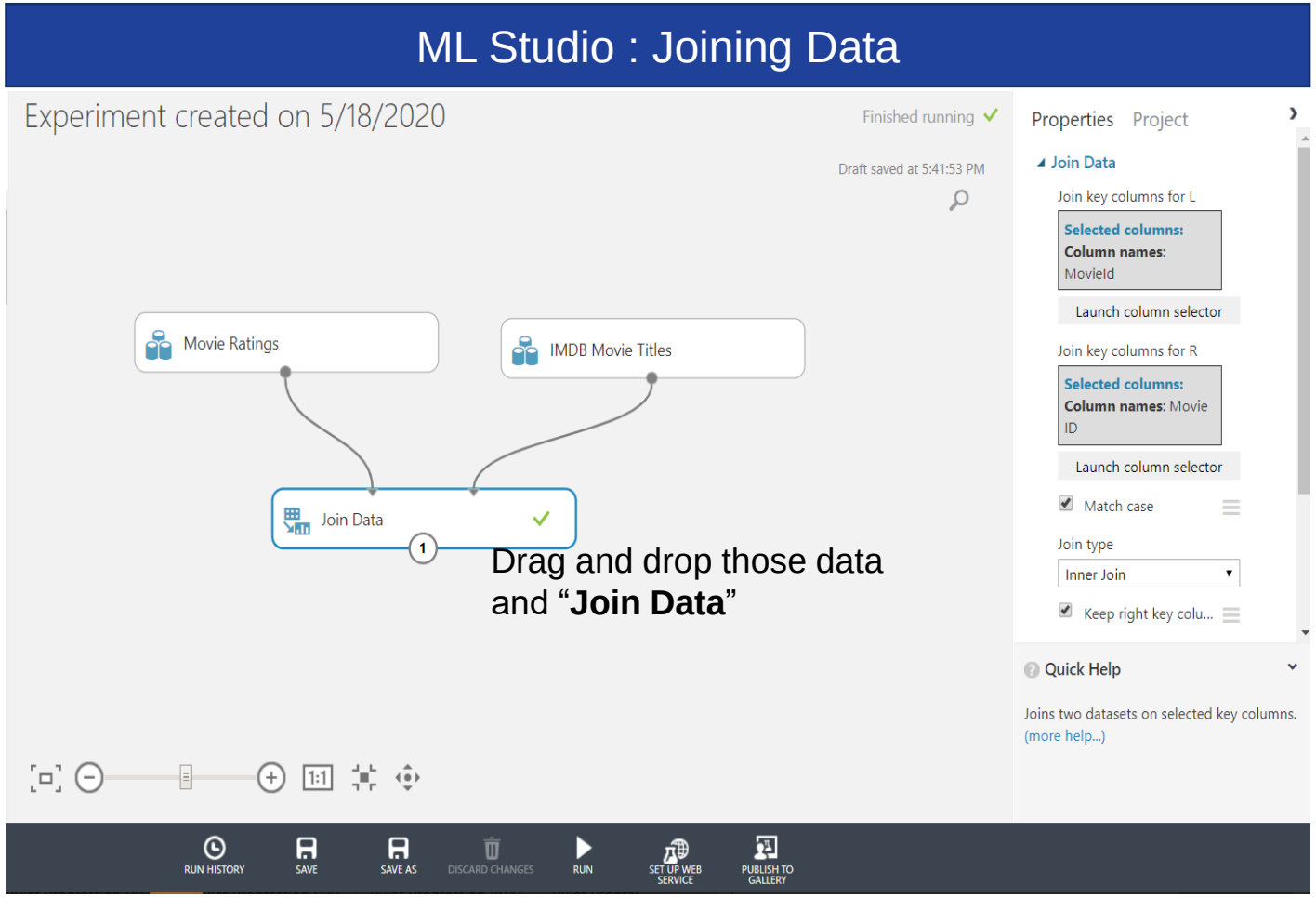
rows 16614 columns 2

view as

Movie ID	Movie Name
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 - lombre de la

Both are original data from two different tables, aims to join this two tables on ‘MovieId’ column.

Joining data - Example



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**

PETRONAS

Joining data - Example

ML Studio : Joining Data- Result

rows 227472 columns 6

	UserId	MovieId	Rating	Timestamp	Movie ID	Movie Name
view as						
	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

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

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/clean-missing-data

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

PETRONAS

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!