



# Data Processing with Azure ML



# Table of Contents

Data Importing Process .....	4
Creation of Experiment .....	8
Entering Data Manually .....	15
Converting File to TSV .....	20
Understanding on Mini View .....	23
Unpacking the Zipped Dataset .....	27
Importing Data from Multiple Sources .....	34
Launch import data wizard .....	38
Data Manipulation Using Add Columns Component.....	40
Data Manipulation Using Add Rows Component .....	45
Data Manipulation Using Remove Duplicate Components .....	49
Data Manipulation Using Select Column in a Dataset Component ...	54
Data Manipulation Using Apply SQL Transformation Component ....	57
Data Manipulation Using Edit Metadata Component .....	63
Data Manipulation Using Clean Missing Data Component .....	69
Data Manipulation Using Partition & Sampling Component .....	72
Data Manipulation Using Split Data Component .....	80

# Goals and Requirements

Estimated time to complete lab is 40-45 minutes.

## Goals

1. Introduce to Azure Machine Learning Environment
2. Azure Machine Learning Data Pre-processing and Cleansing.
3. Implementing Data Transformations.

## Requirements

1. Access to an Azure Machine Learning Studio and related documents provided in the session.

# Data Input Output to Azure Workspace

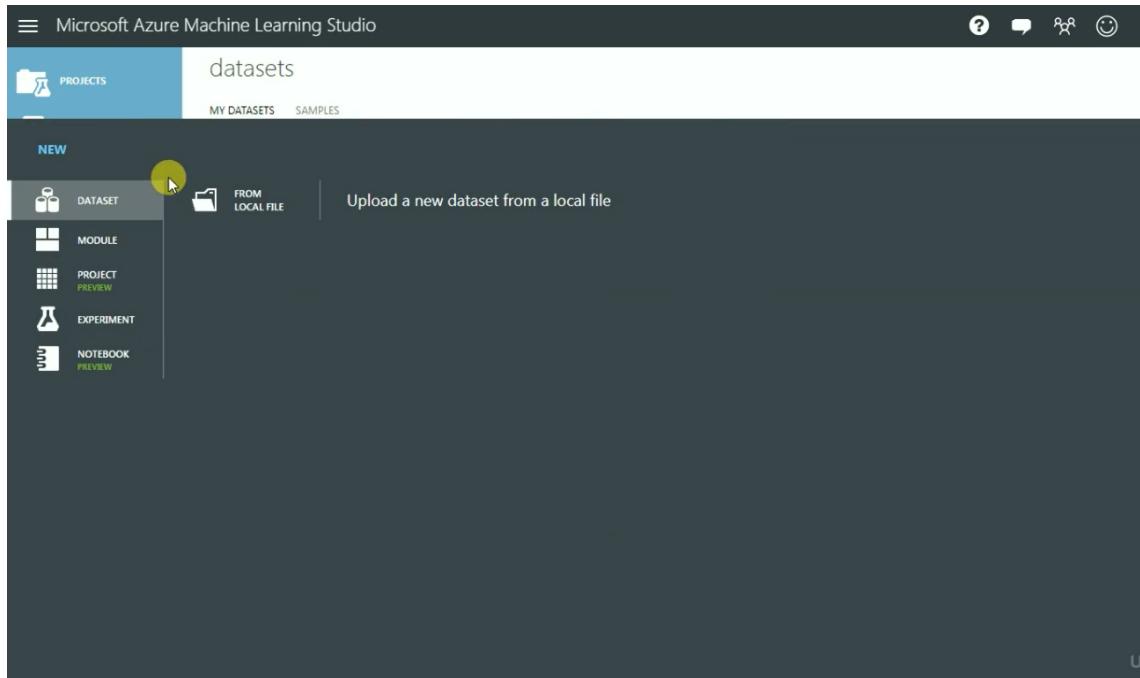
## Data Importing Process

The screenshot shows the Microsoft Azure Machine Learning Studio interface. On the left, there is a sidebar with icons for Projects, Experiments, Web Services, Notebooks, Datasets (selected), Trained Models, and Settings. The main area is titled "datasets" and contains a table titled "MY DATASETS". The table has columns for NAME, SUBMITTED BY, DESCRIPTION, DATA TYPE, CREATED, SIZE, and PROJECT. One row is visible: "Churn\_Modelling.csv" submitted by "My New ML experiment for..." with a description "GenericCSV", created on "1/20/2017 3:14:31 PM", size "668.81 KB", and project "None". At the bottom of the table are buttons for DOWNLOAD, DELETE, OPEN IN NOTEBOOK, GENERATE DATA ACCESS CODE..., and ADD TO PROJECT. A yellow circle highlights the "NEW" button at the bottom left of the table.

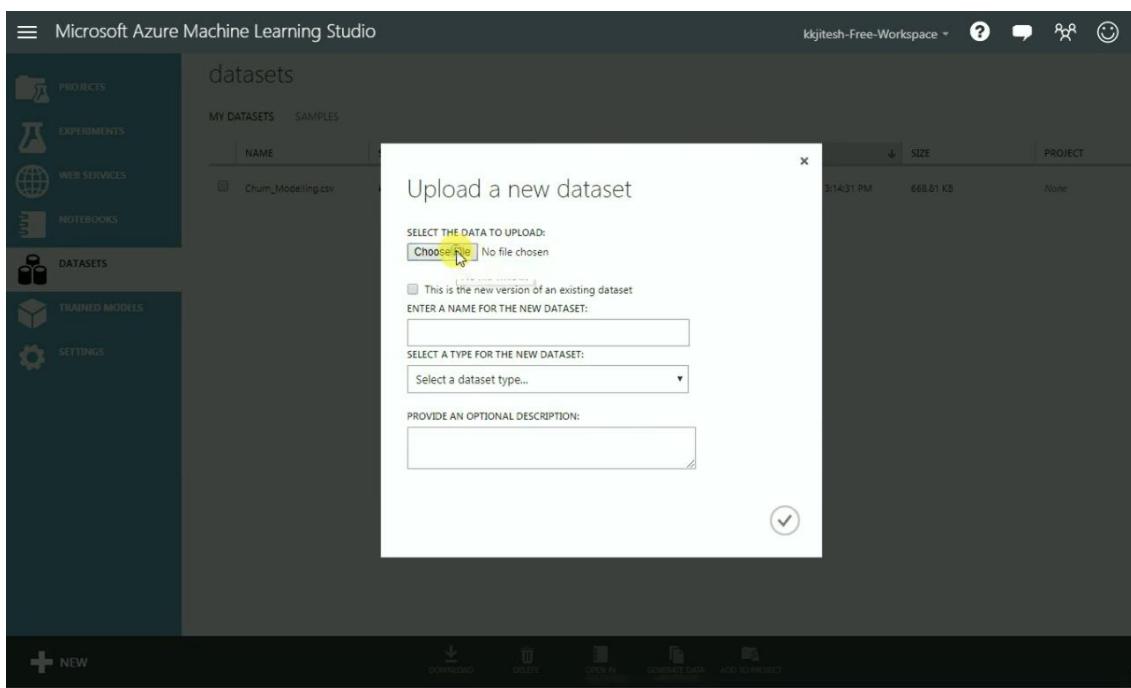
Click on New

This screenshot is identical to the one above, showing the Microsoft Azure Machine Learning Studio interface with the "datasets" page. A yellow circle highlights the "NEW" button at the bottom left of the "MY DATASETS" table, indicating where the user should click to start a new dataset entry.

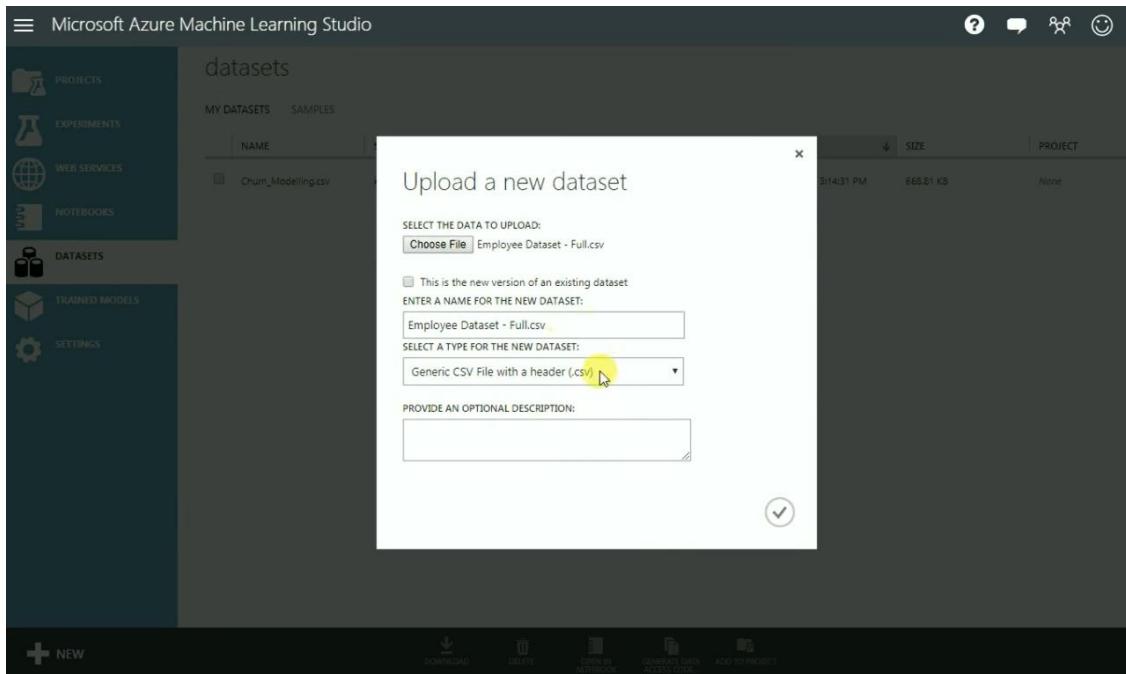
Go to dataset and click on from Local File



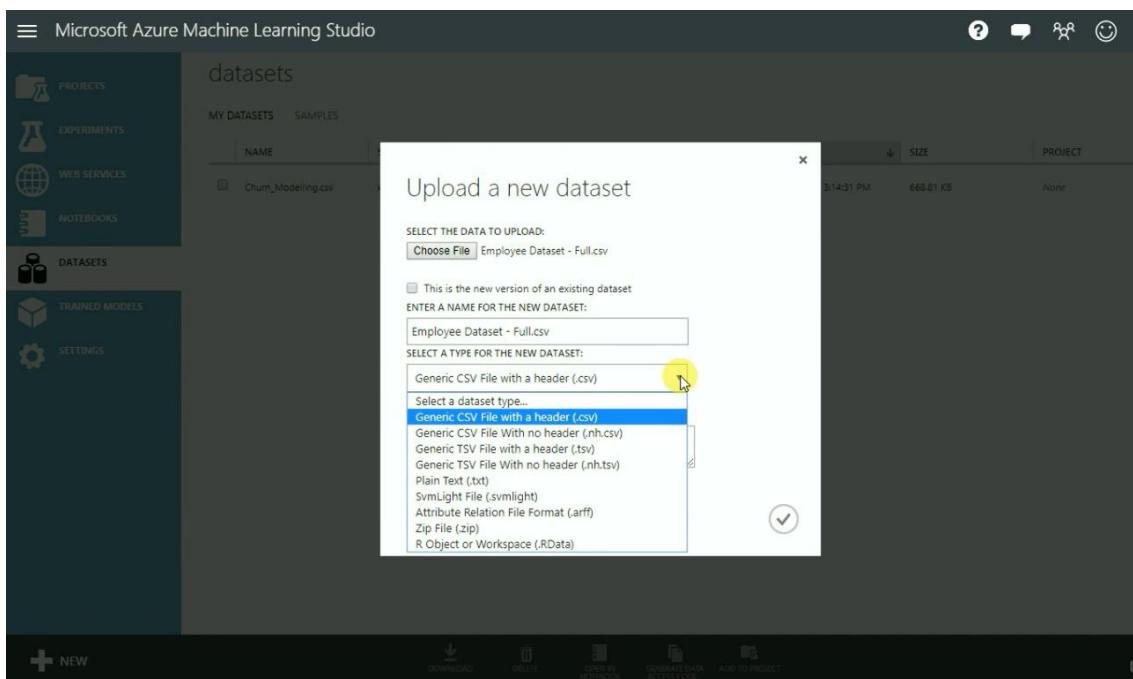
Click on choose a file



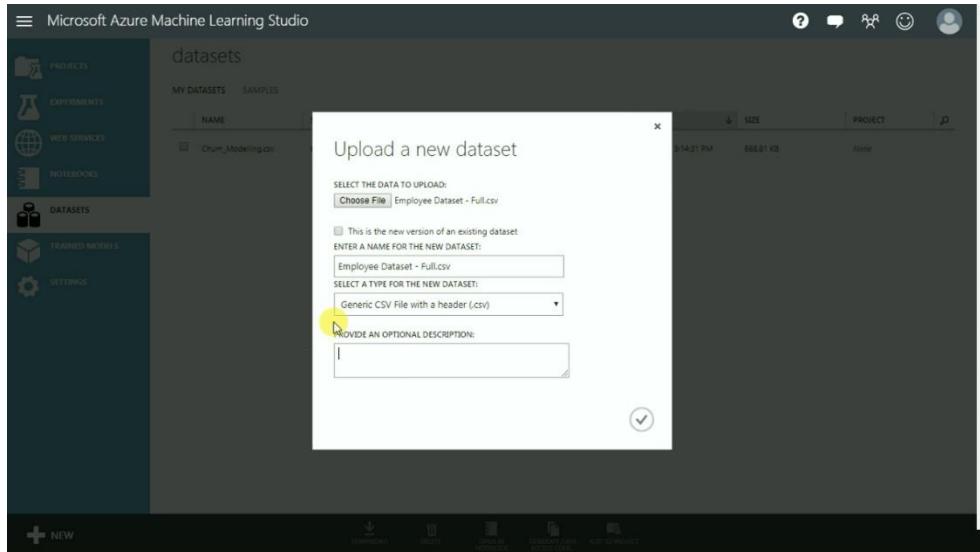
Select the dataset Employee-Dataset-Full.csv from the local folder shared.



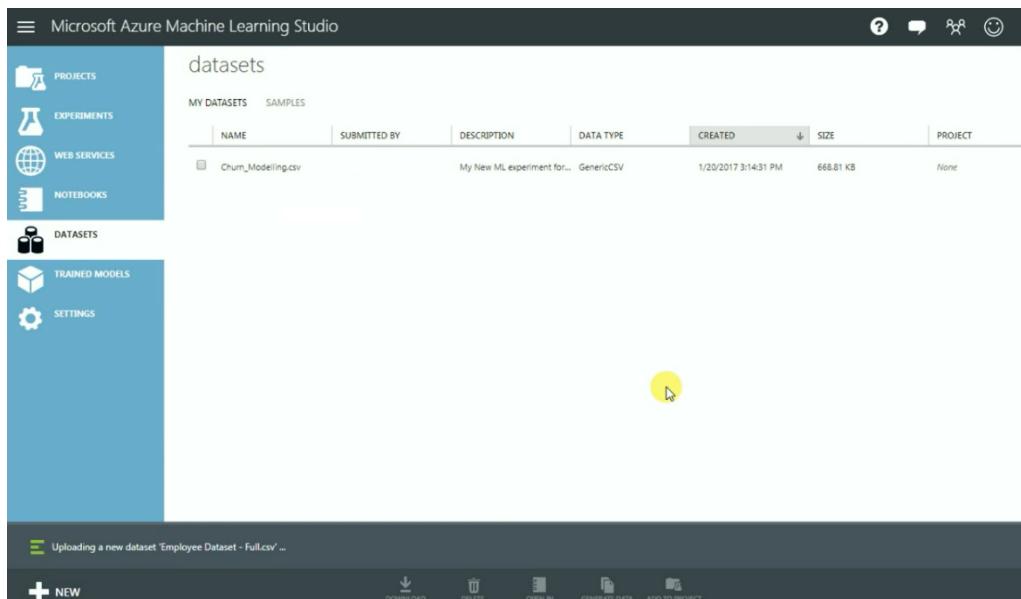
Can select any options from the dropdown menu



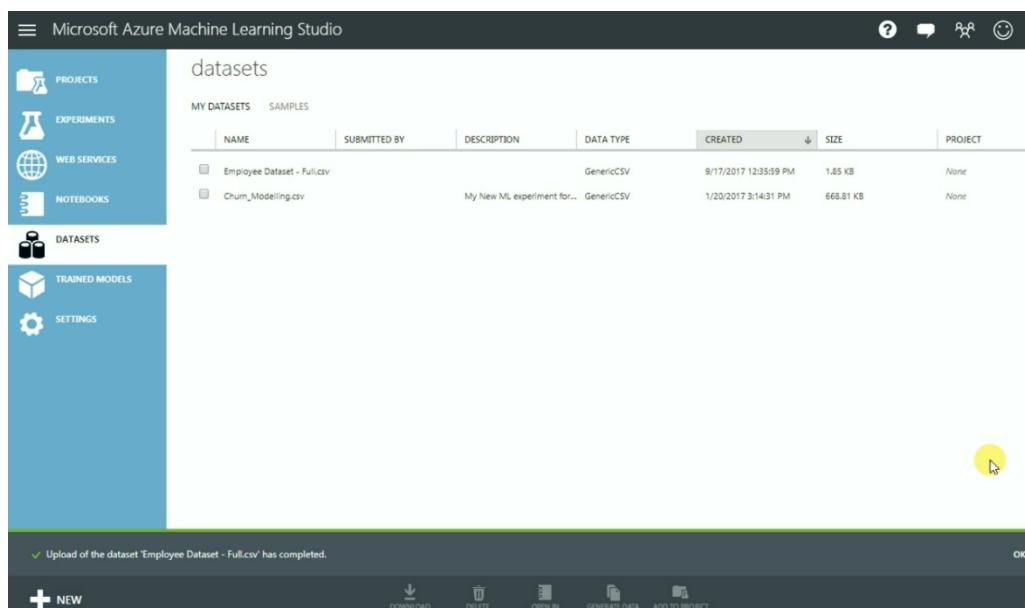
Can provide additional description also as shown below and then click ok



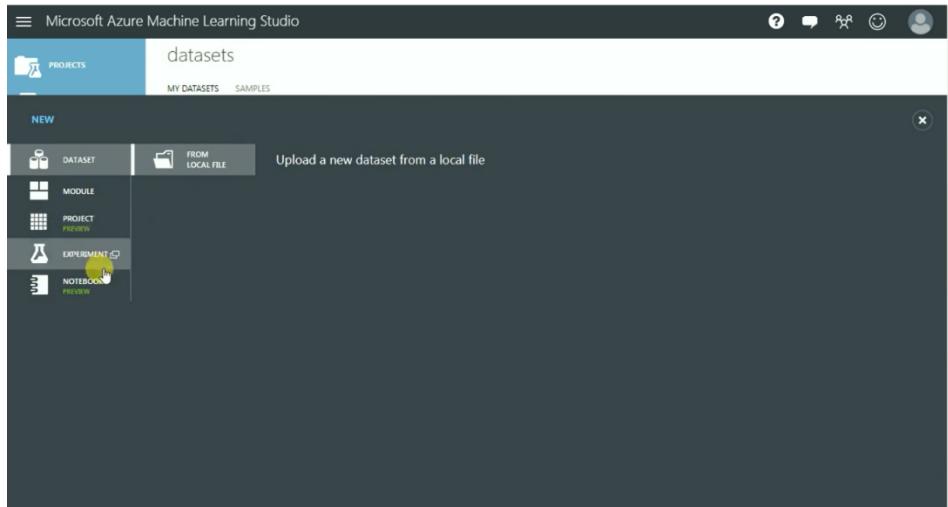
Wait for dataset update



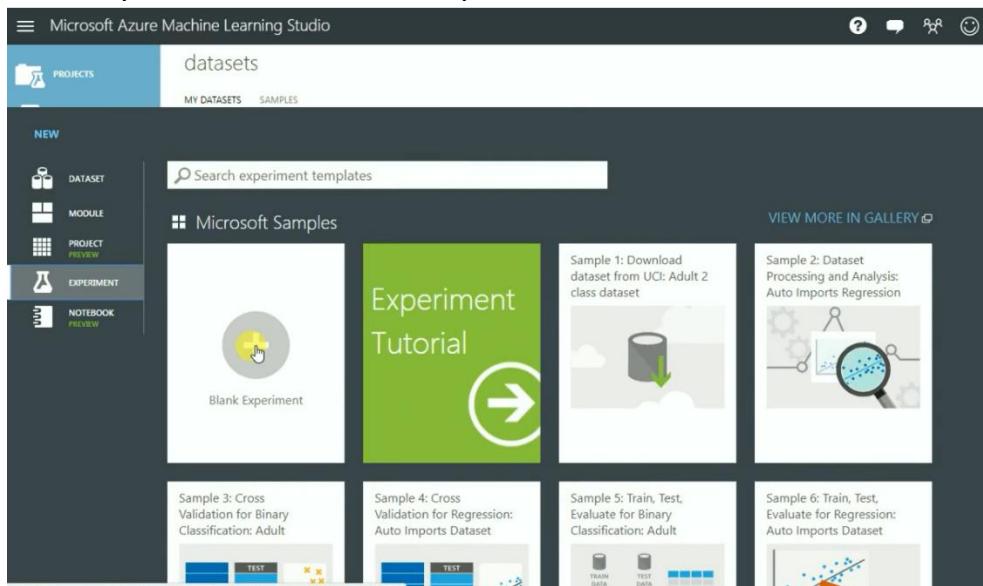
After updating click ok and then click new



## Creation of Experiment



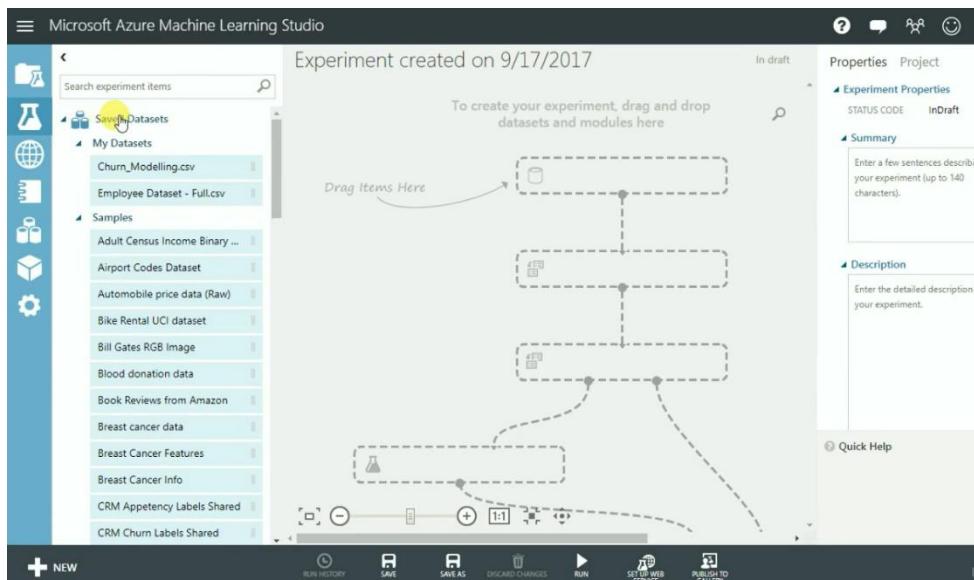
Click experiment and blank experiment



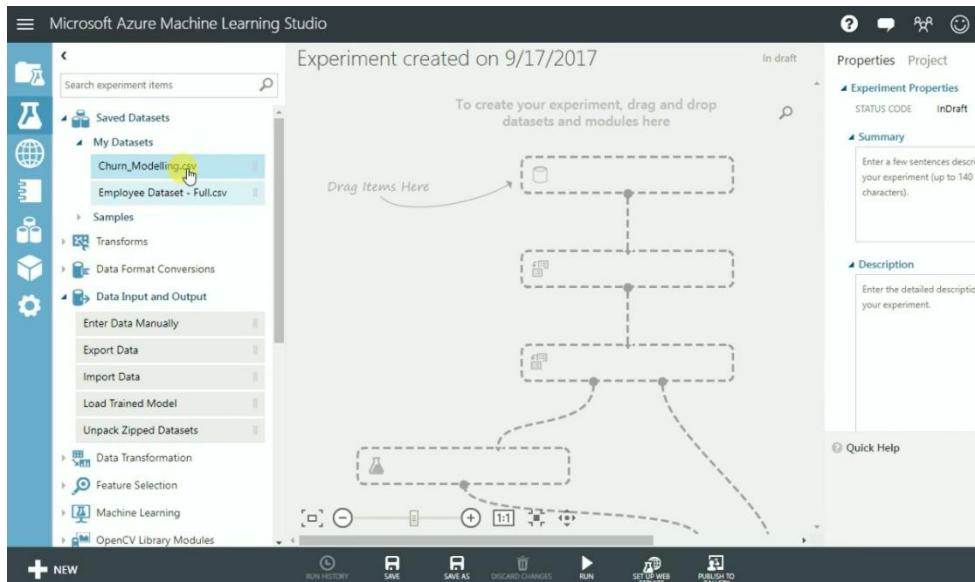
Check the datasets



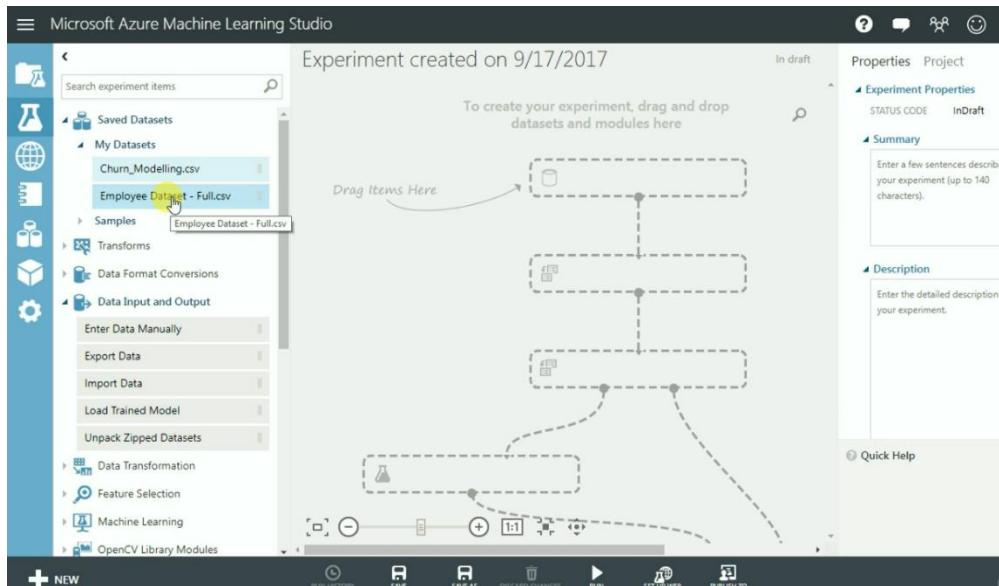
Click on saved data sets



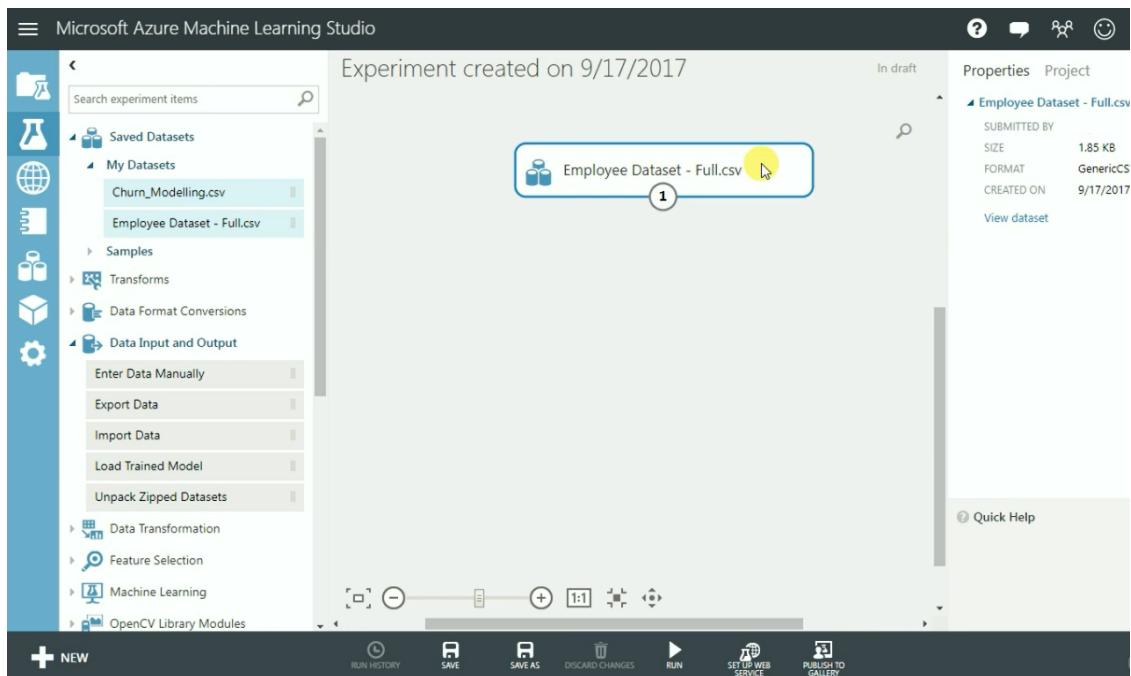
Dataset updated



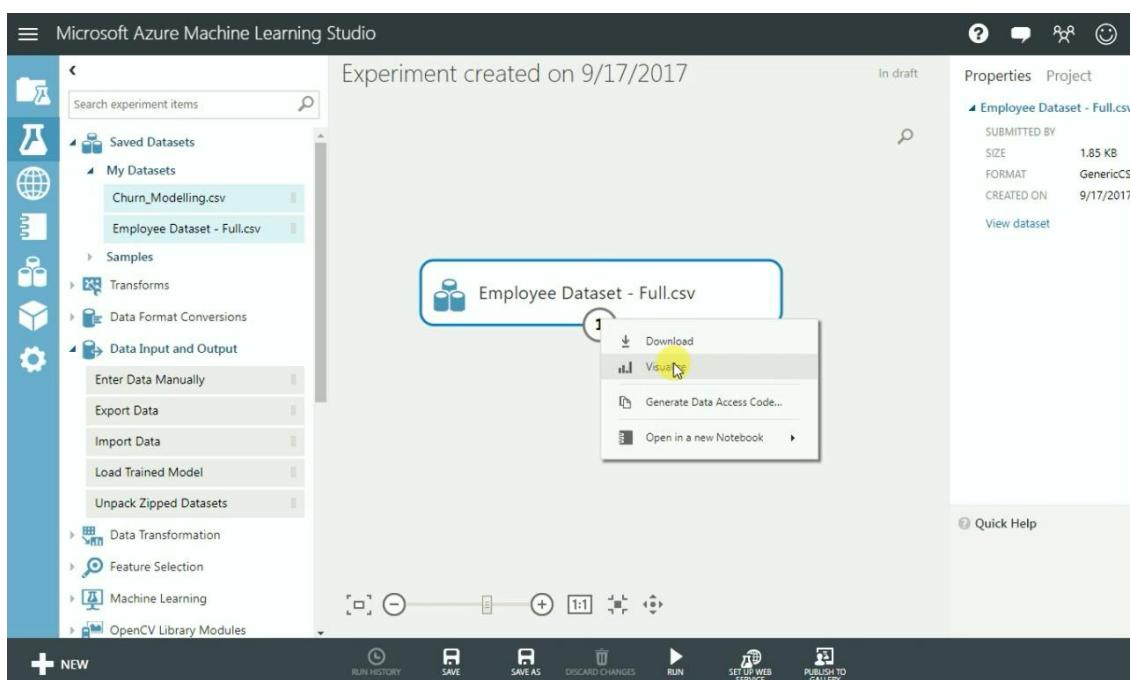
Post updating drag and drop the data set



You can find flowchart structured rectangle with a node



Right Click on node 1 and then click visualize



Microsoft Azure Machine Learning Studio

Experiment created on 9/17/2017

rows 25 columns 11

Employee Name	Age	Last Working Day	Department	Education	Gender	Marital Status
Jitesh	41	31-12-9999	Training	Masters	Male	Sing
Sanjit	49	31-12-9999	Sales	Masters	Male	Mar
John	37	31-12-9999	R&D	Doctorate	Male	Sing
Sandra	33	31-12-9999	Software Development	Undergraduate	Female	Mar
Madhu	27	31-12-9999	R&D	Masters	Male	Mar
Robert	32	31-12-9999	R&D	Masters	Male	Sing
Megan	59	31-12-9999	Software Development	Masters	Female	Mar
Matt	30	31-12-9999	R&D	Doctorate	Male	Div

view as

To view, select a column in the table.

Statistics

Visualizations

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

Can also view the graph

Microsoft Azure Machine Learning Studio

Experiment created on 9/17/2017

rows 25 columns 11

Last Working Day	Department	Education	Gender	Marital Status	Monthly Income	Years of Experience	Percent Salary Hike
31-12-9999	Training	Masters	Male	Single	5993	8	11
31-12-9999	Sales	Masters	Male	Married	5130	1	23
31-12-9999	R&D	Doctorate	Male	Single	2090	6	15
31-12-9999	Software Development	Undergraduate	Female	Married	2909	1	11
31-12-9999	R&D	Masters	Male	Married	3468	9	12
31-12-9999	R&D	Masters	Male	Single	3068	0	13
31-12-9999	Software Development	Masters	Female	Married	2670	4	20
31-12-9999	R&D	Doctorate	Male	Divorced	2693	1	22

To view, select a column in the table.

Statistics

Visualizations

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

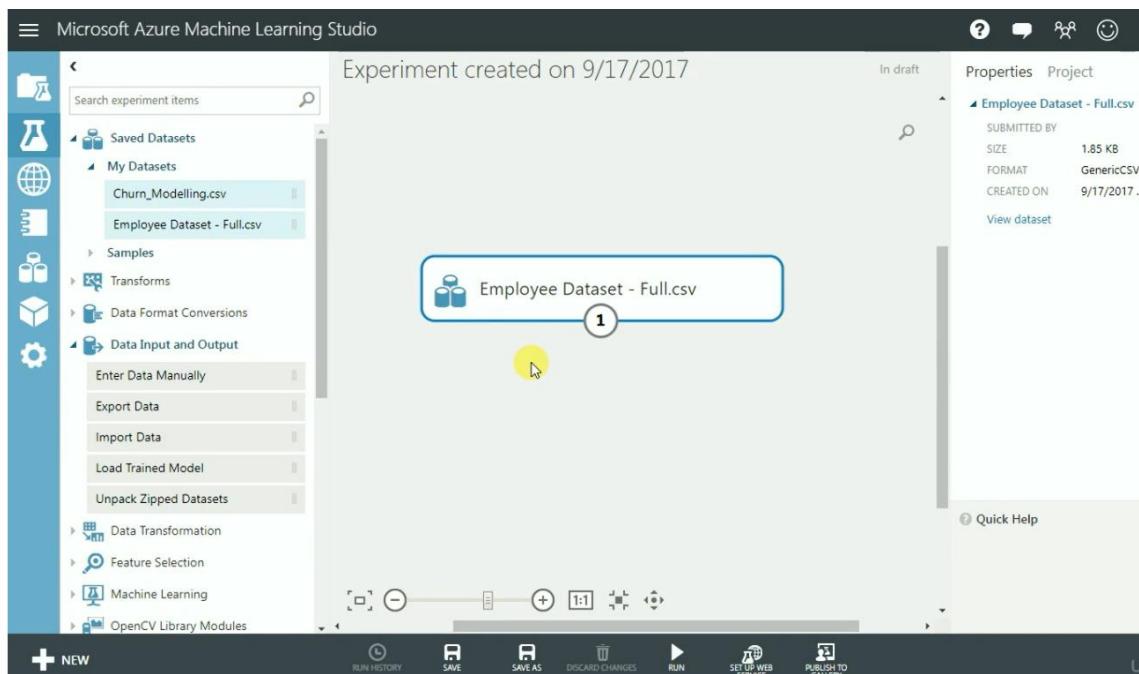
You can view statistics and graph if any data selected

The screenshot shows the Microsoft Azure Machine Learning Studio interface. On the left, there's a sidebar with various icons for file operations like New, Save, and Publish. The main area displays a dataset titled "Employee Dataset - Full.csv" with 25 rows and 11 columns. The columns include "Last Working Day", "Department", "Education", "Gender", "Marital Status", "Monthly Income", "Years of Experience", and "Percent Salary Hike". To the right of the table, there are two sections: "Statistics" and "Visualizations". The "Statistics" section shows that there are 2 unique values, 1 missing value, and the feature type is a String Feature. The "Visualizations" section shows a histogram for the "Gender" column, comparing it to "None". The histogram has two bars: one for "Male" reaching a frequency of 16, and one for "Female" reaching a frequency of approximately 8.

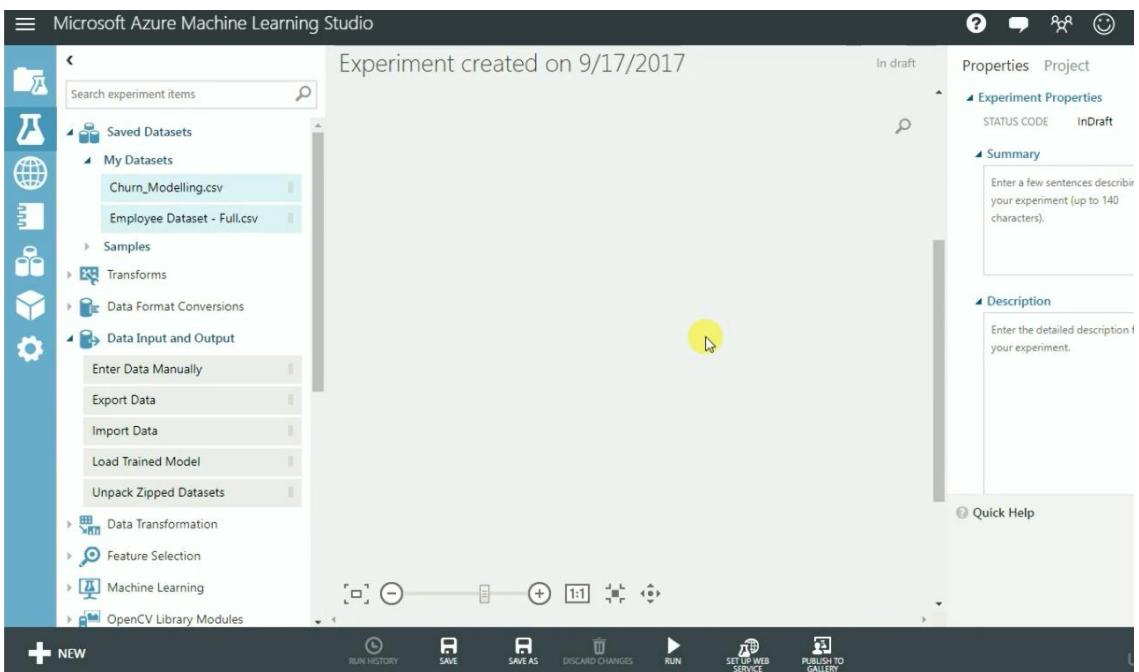
You can compare the data from dropdown

This screenshot is similar to the previous one but focuses on the "compare to" dropdown in the "Visualizations" section. The dropdown menu is open, showing options like "None", "Employee Name", "Age", "Last Working Day", "Department", "Education", "Gender", "Marital Status", "Monthly Income", "Years of Experience", "Percent Salary Hike", and "Performance Rating". The "None" option is currently selected. The histogram for the "Gender" column is visible in the background, showing the same data as the first screenshot.

Close the visualization and go back to input

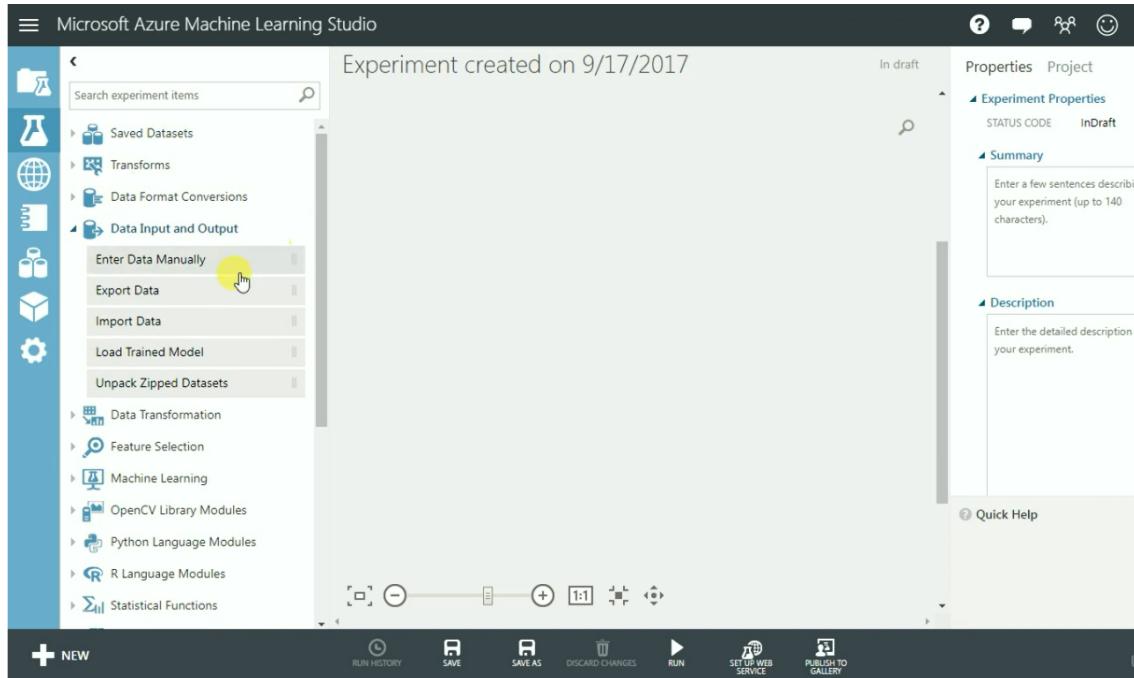


Delete the dataset and have some free space

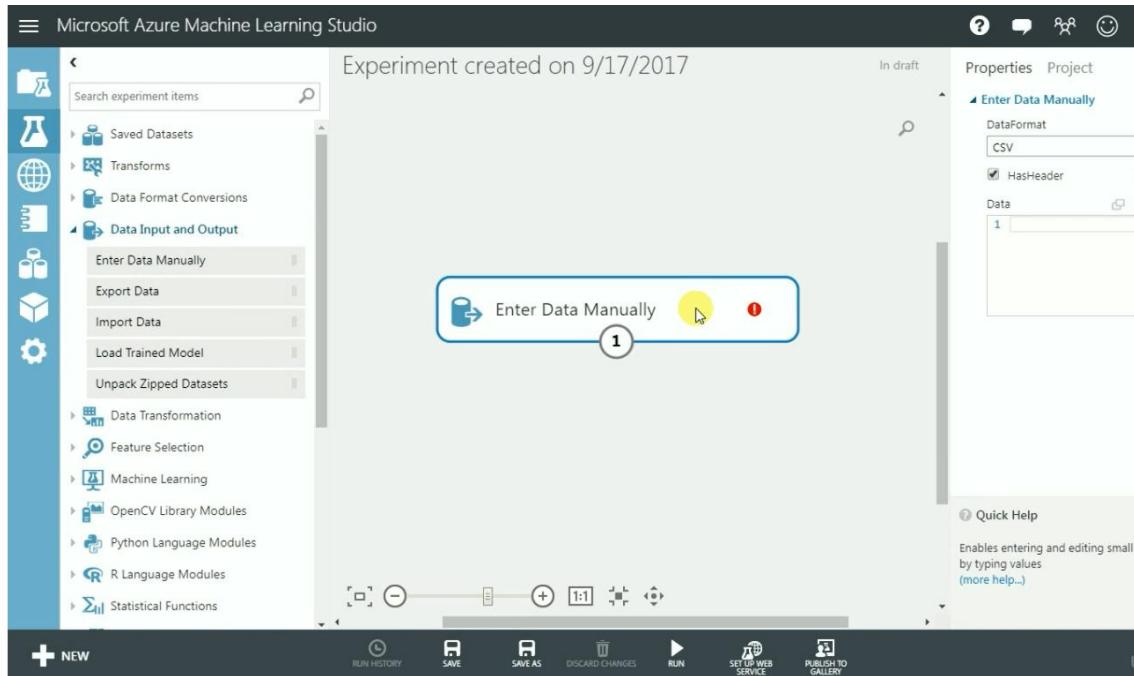


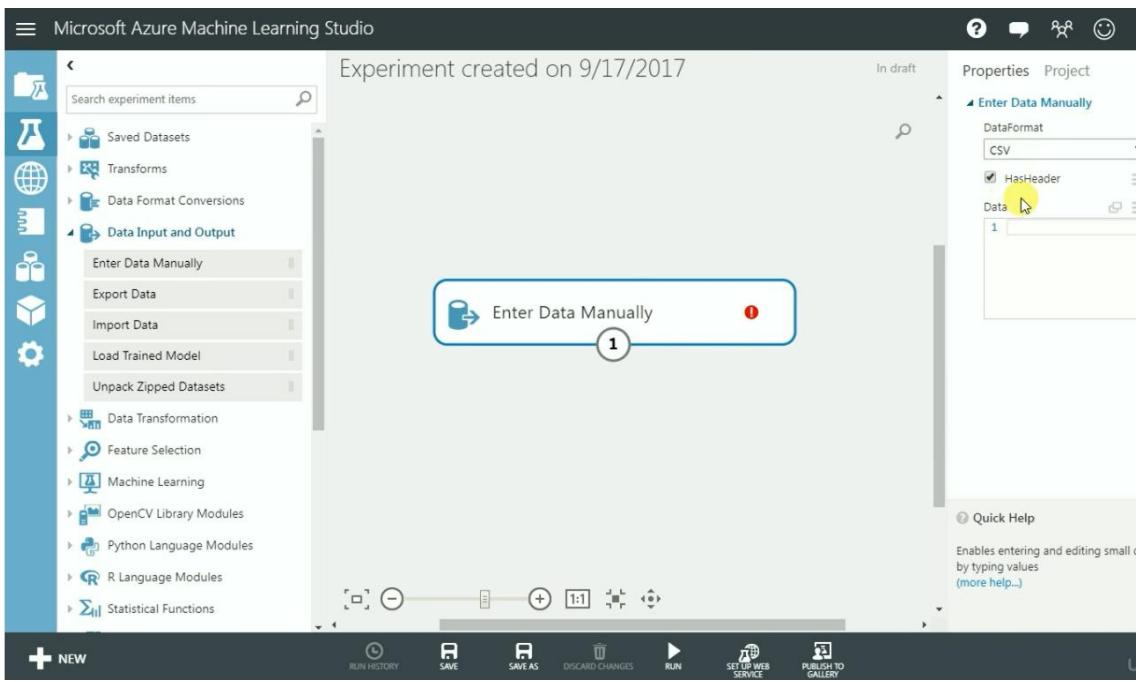
## Entering Data Manually

Now we can see how to enter manual data

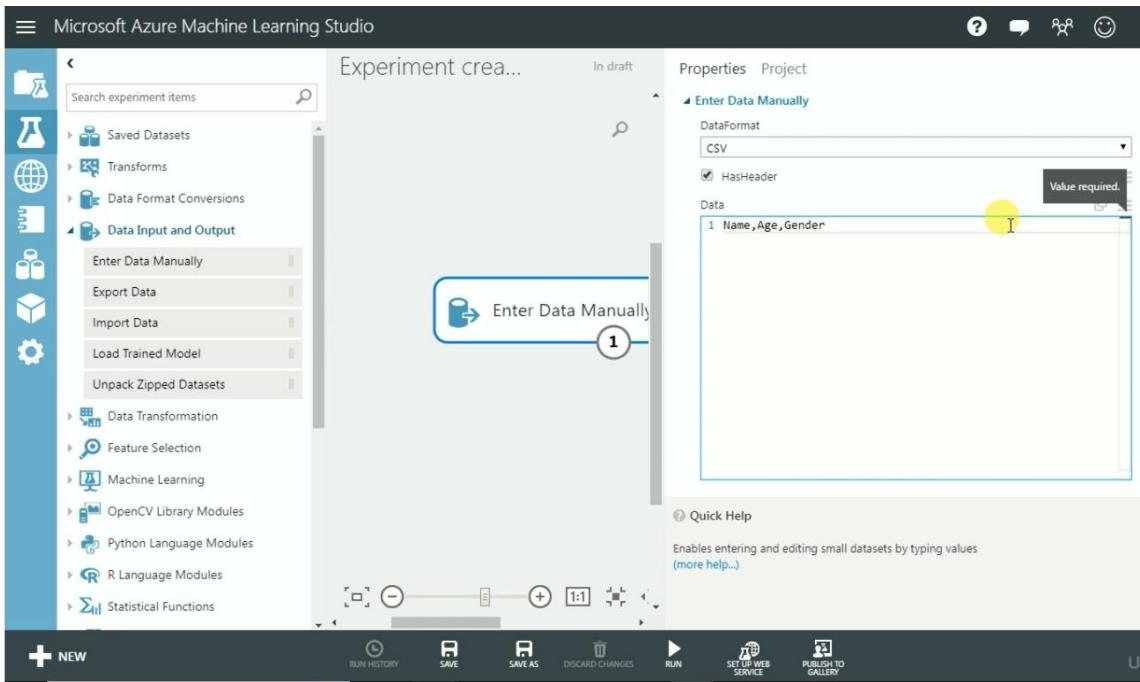


Drag and drop Enter data manually

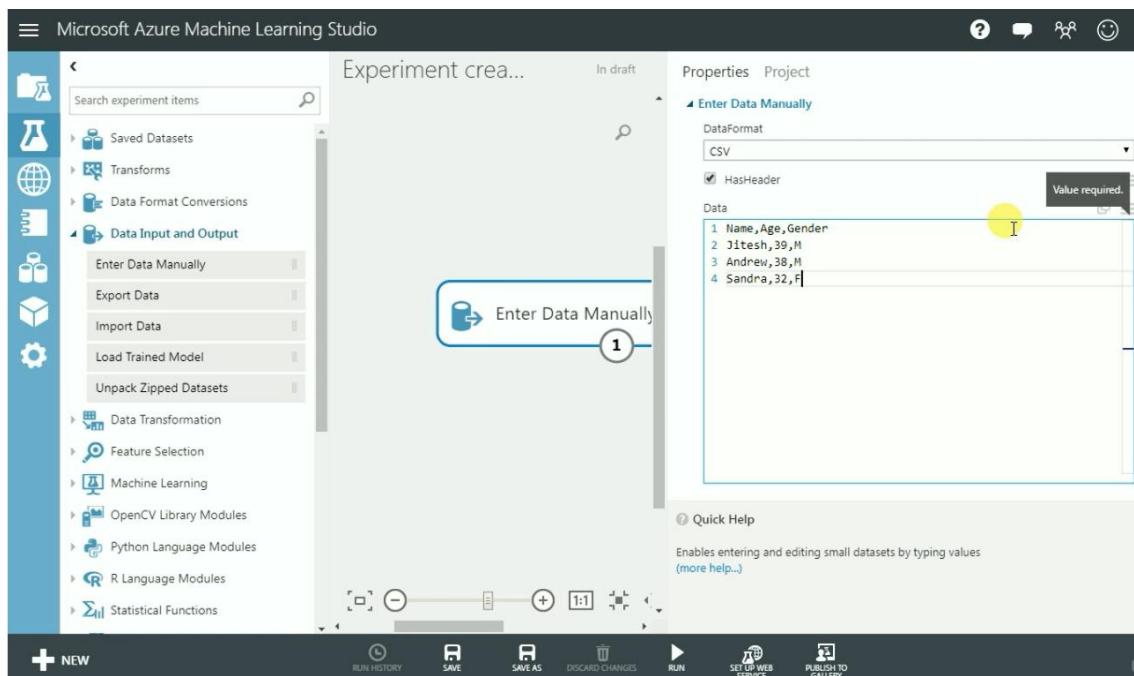




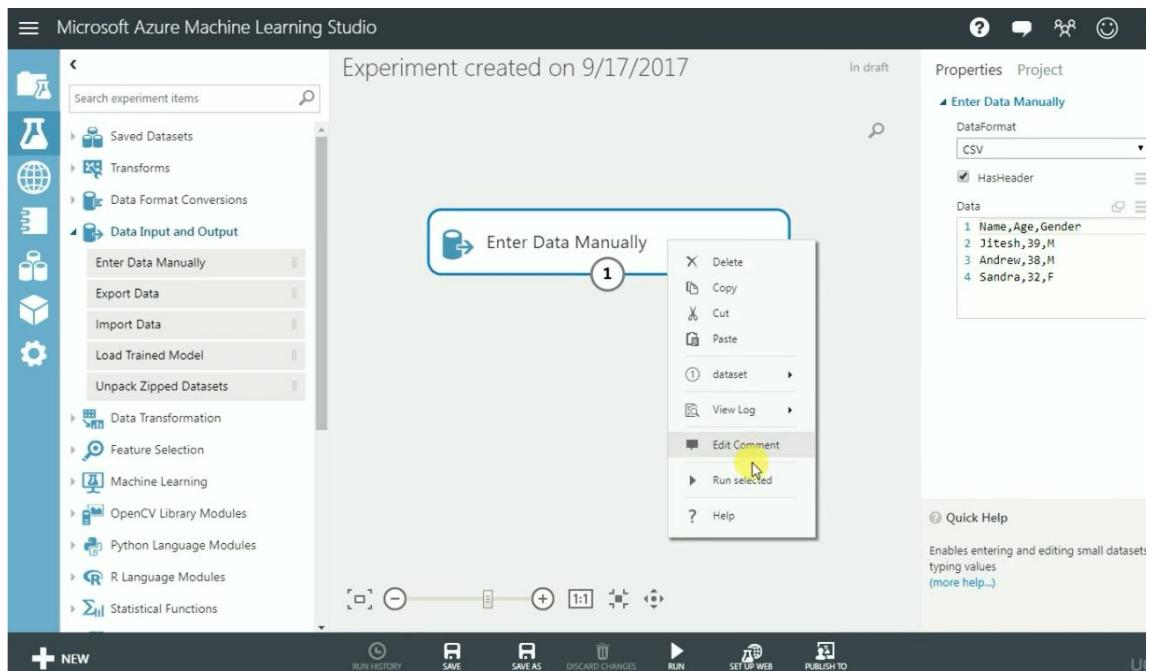
Type heading as shown below



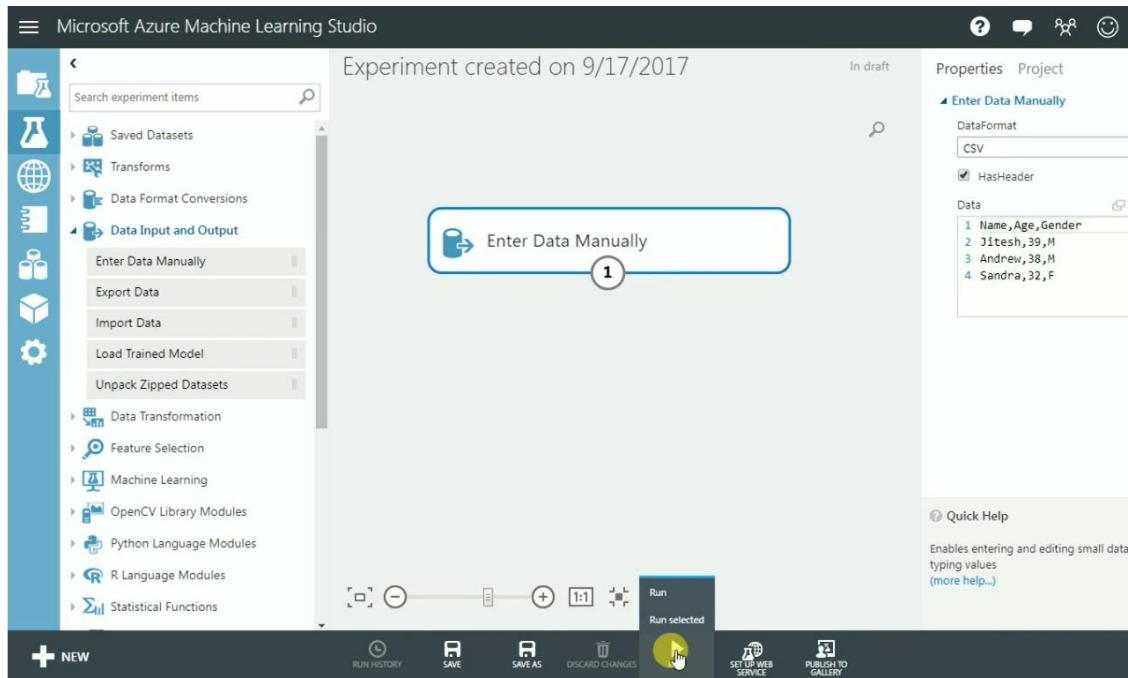
Type the data one by one



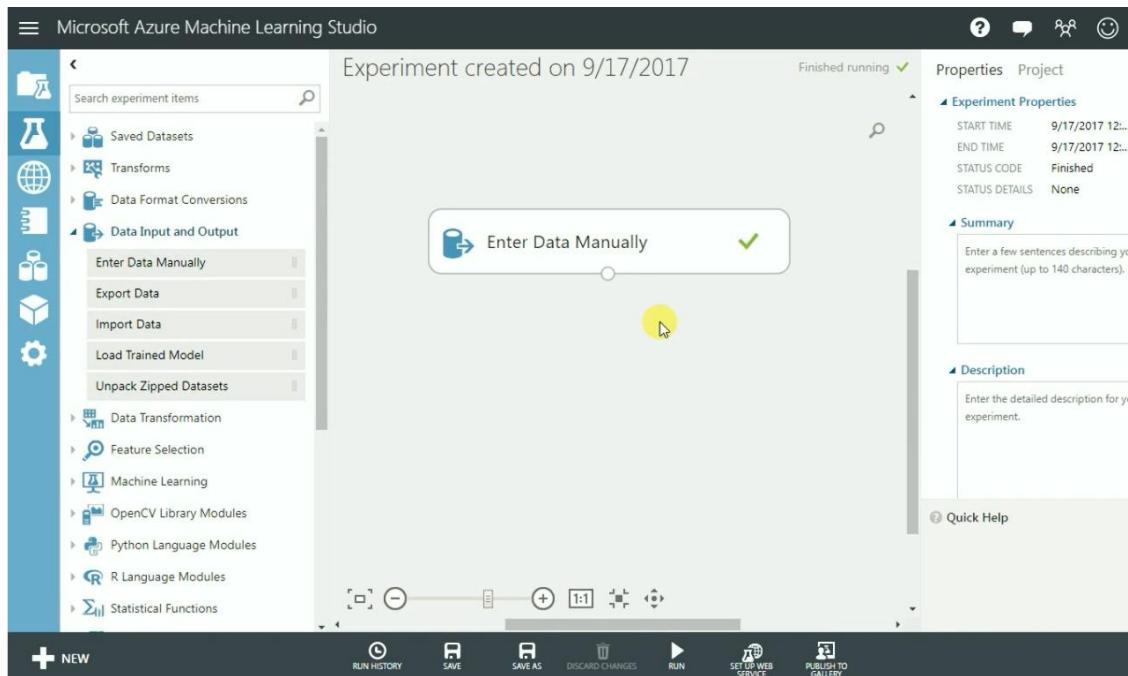
To run- right click and click run selected



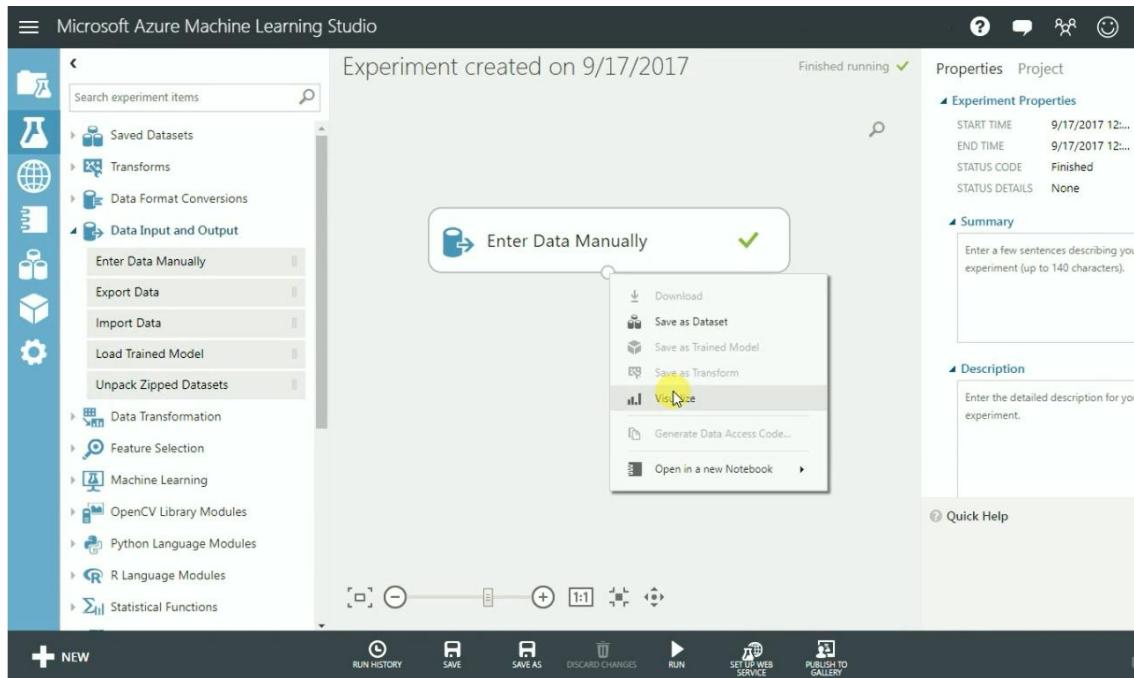
Or simply click run in the bottom



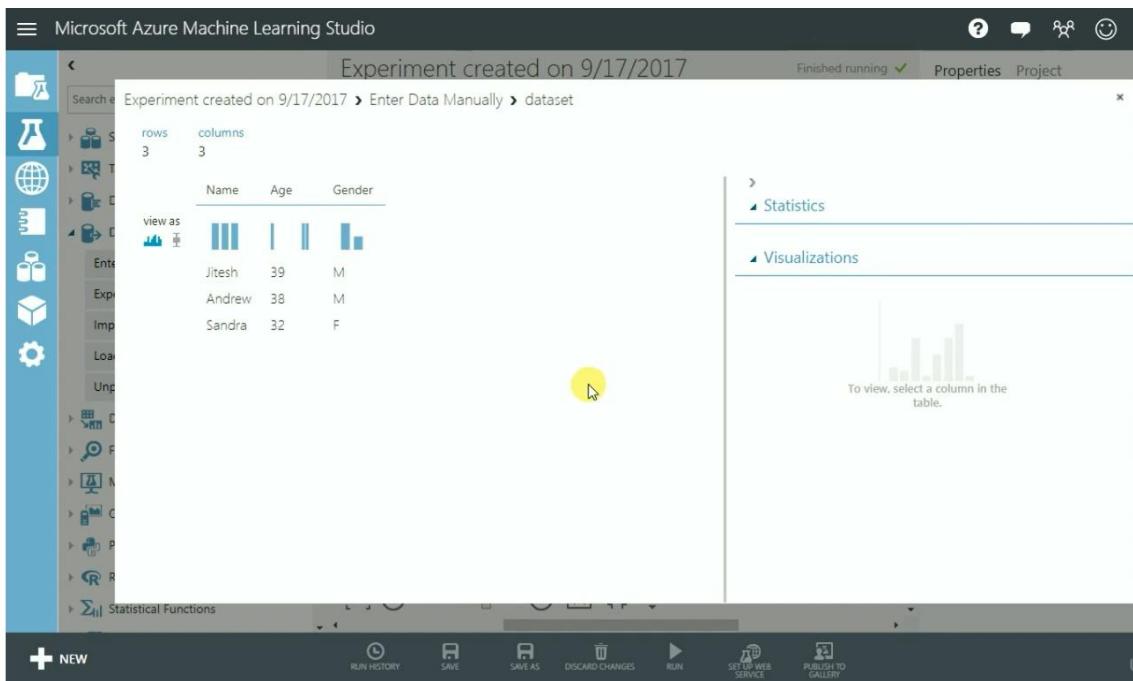
Finished running



## Right click and visualize

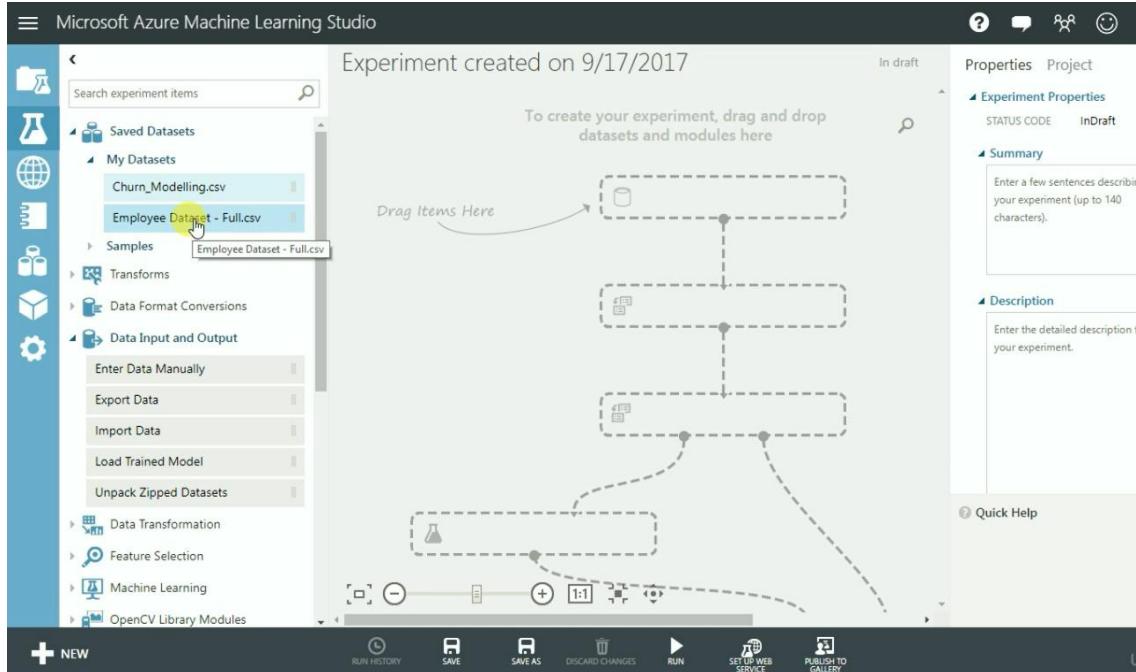


Created dataset with specified inputs successfully

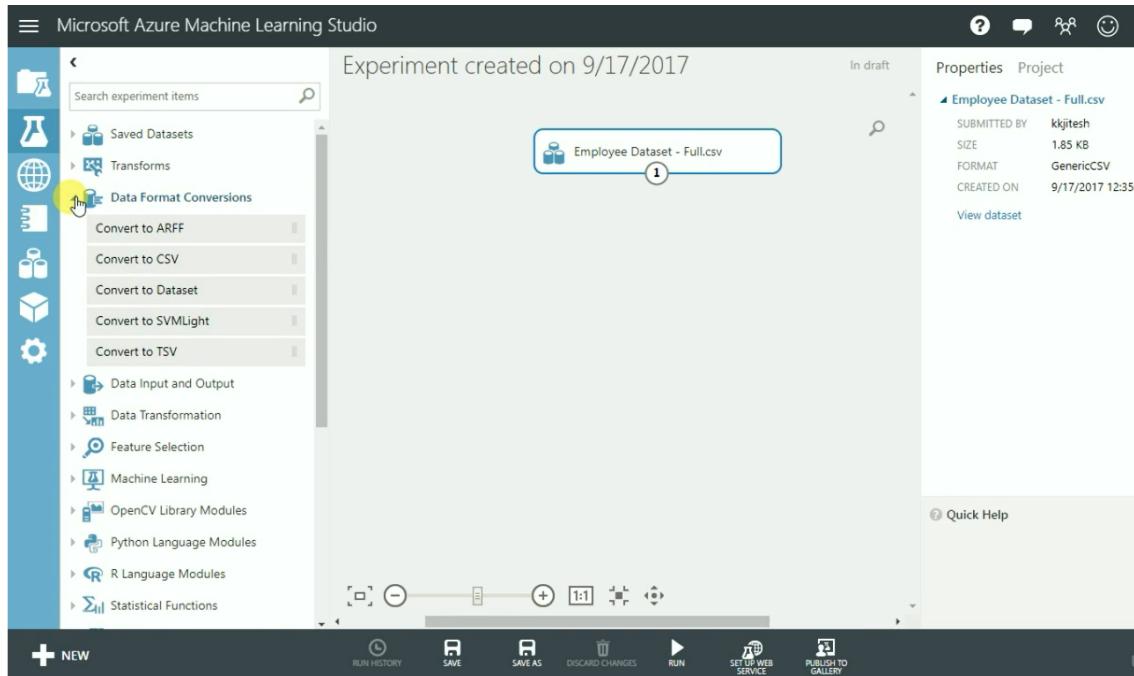


## Converting File to TSV

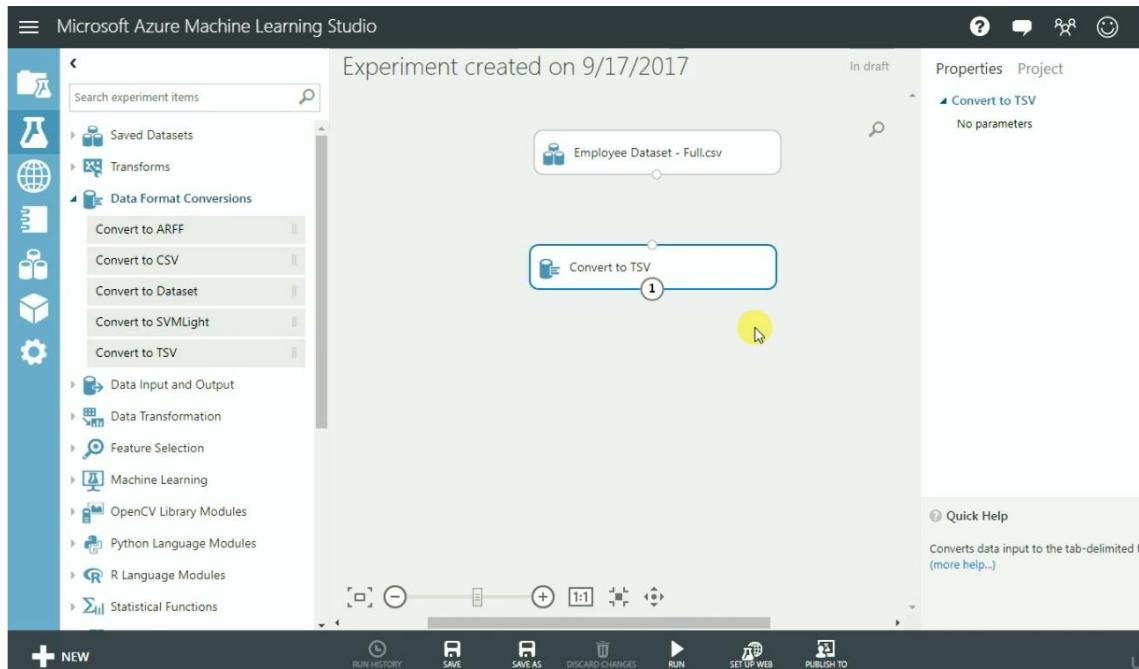
Pick dataset created earlier



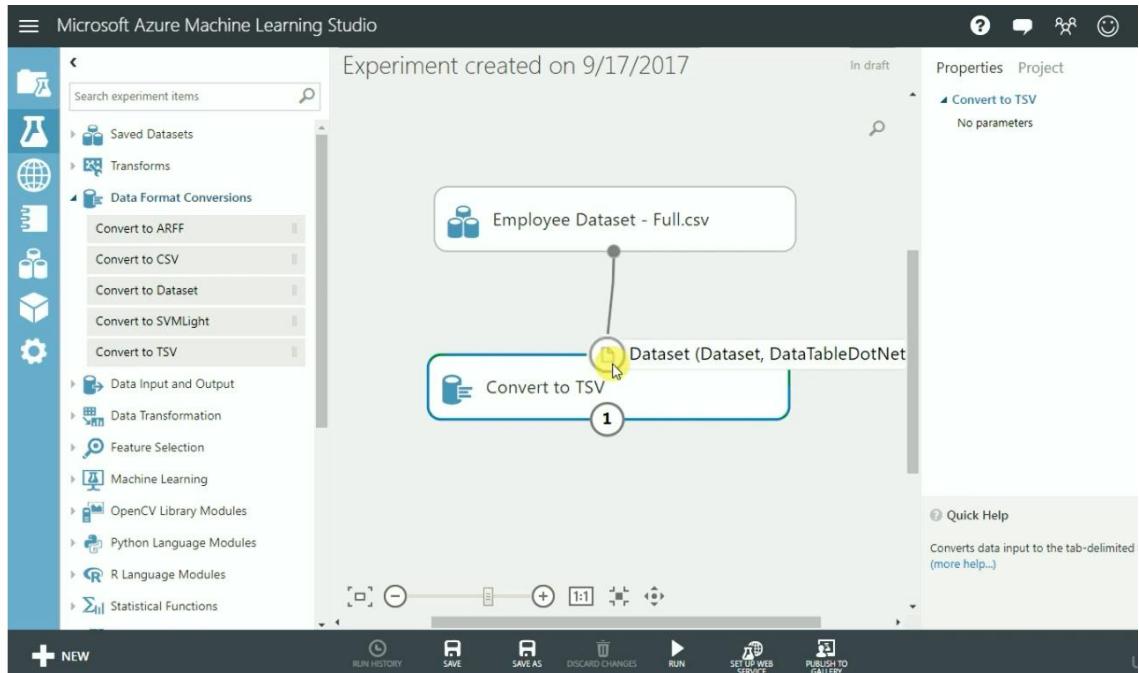
Drag and drop the dataset and click on data format conversion



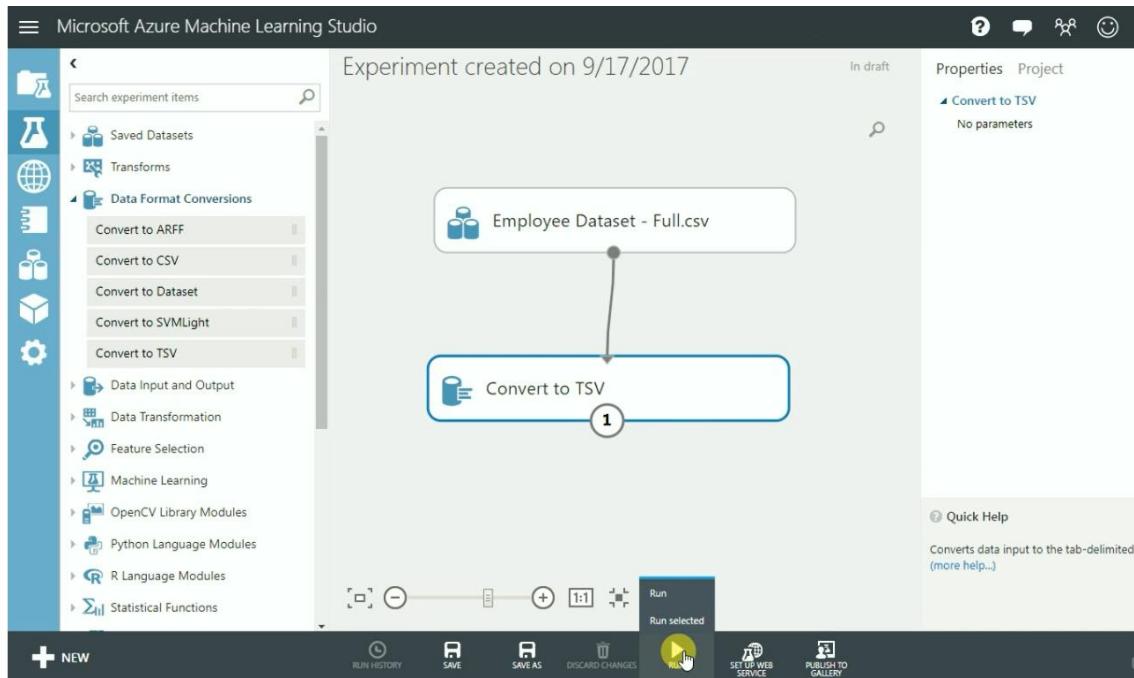
## Drag and drop convert to TSV module



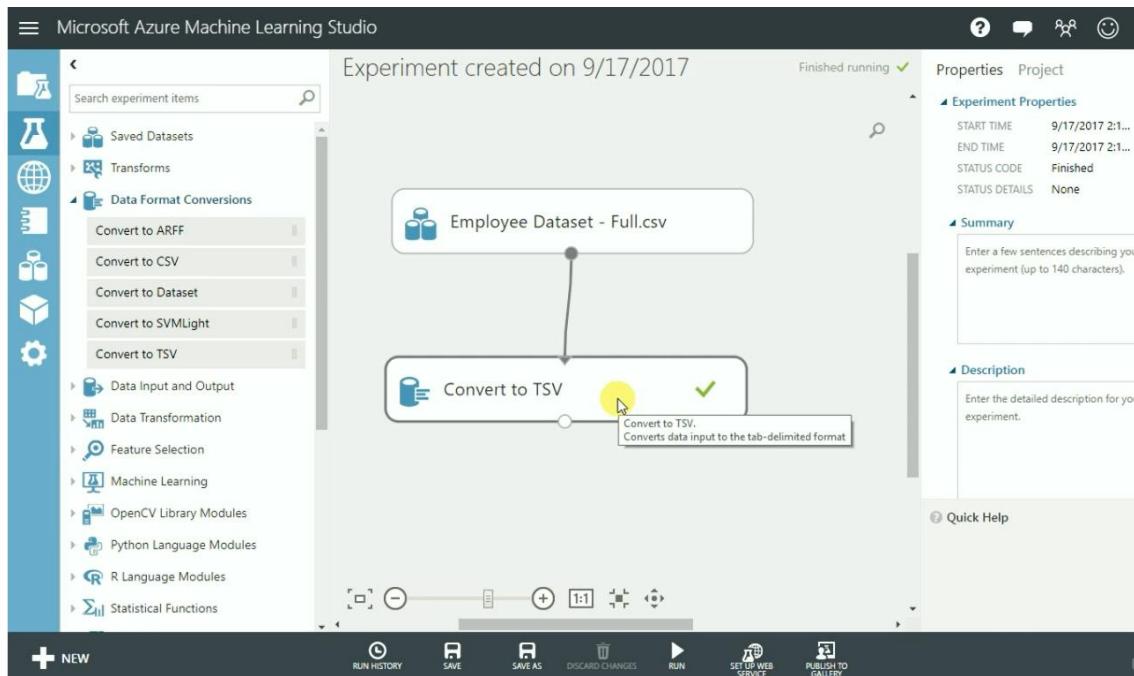
Simply attach Output node from CSV to TSV as shown



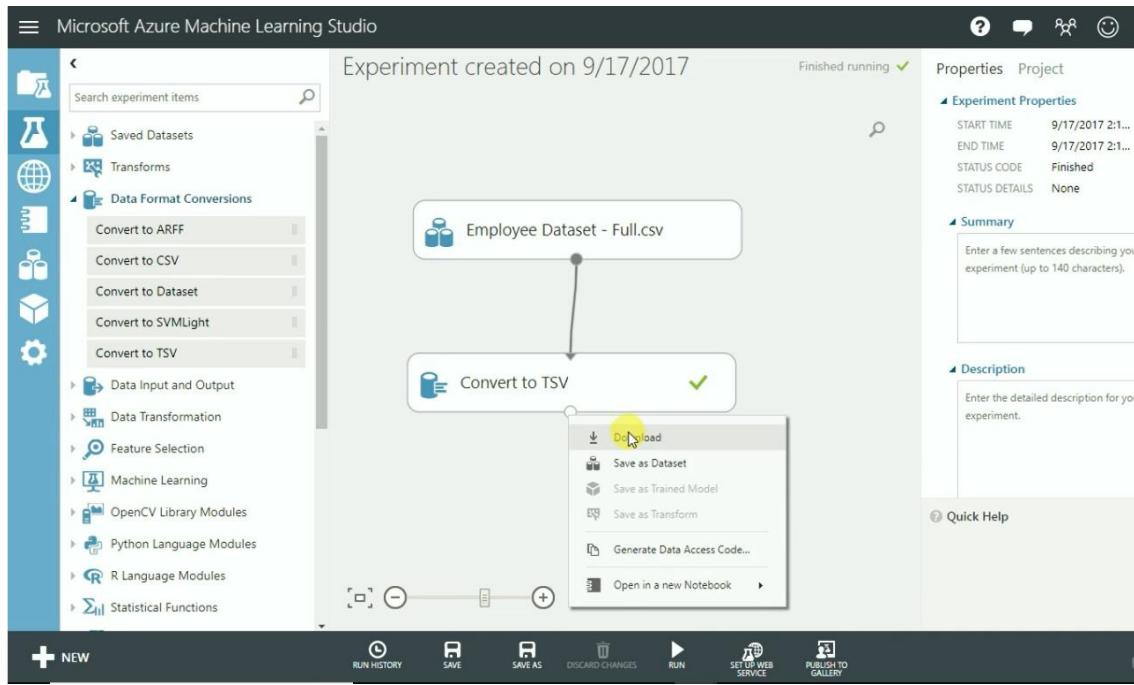
## Run the module for execution



Now Azure has converted from csv to tsv file.

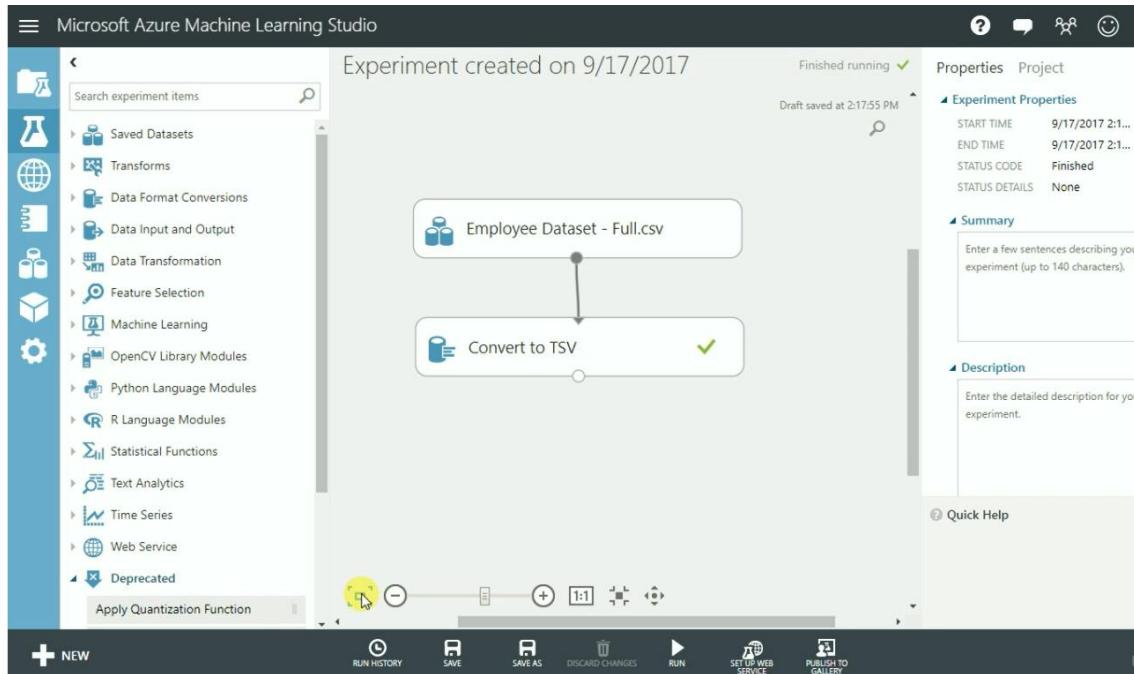


Can download the same by right click and download to check the output

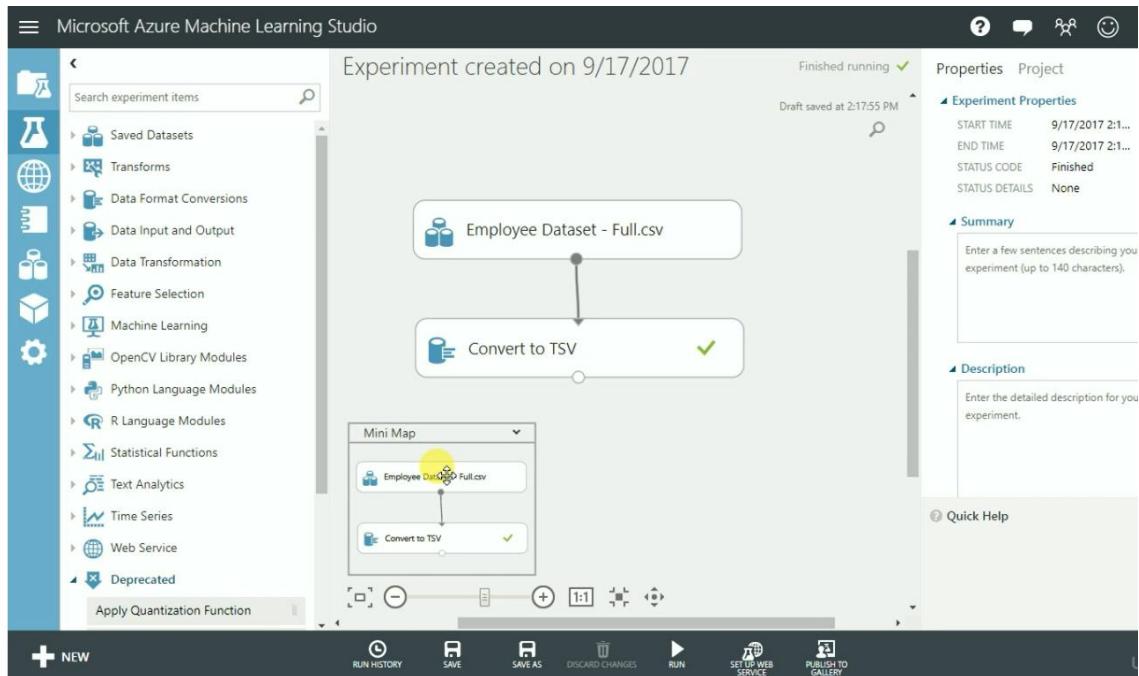


## Understanding on Mini View

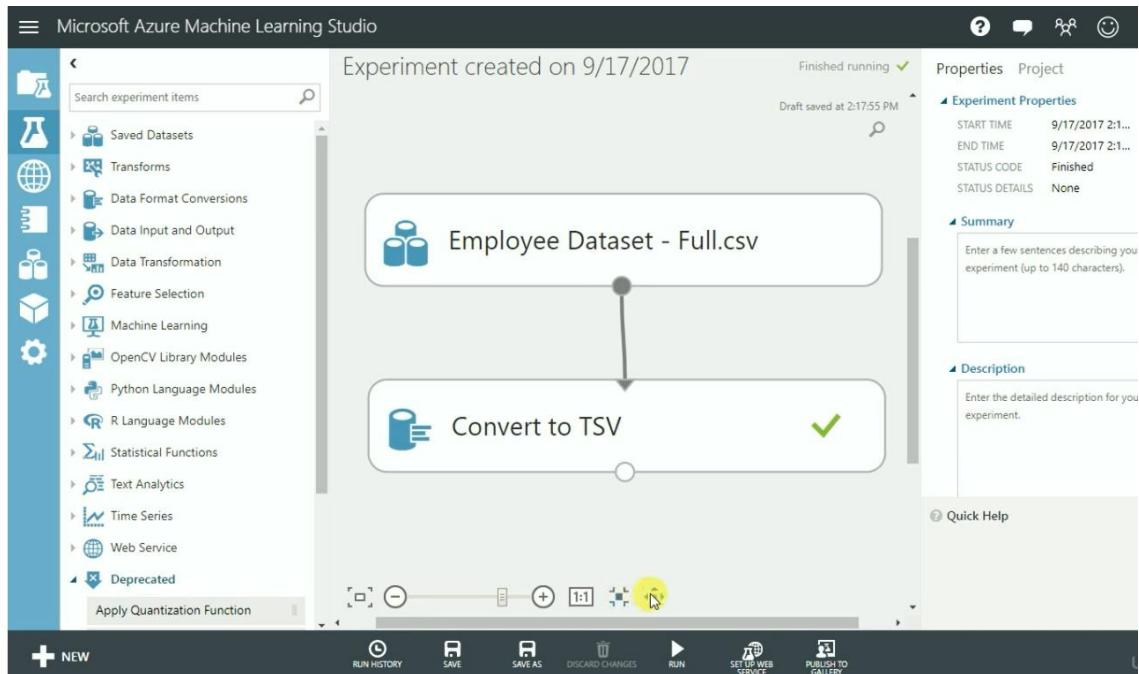
Click on mini map to preview the module



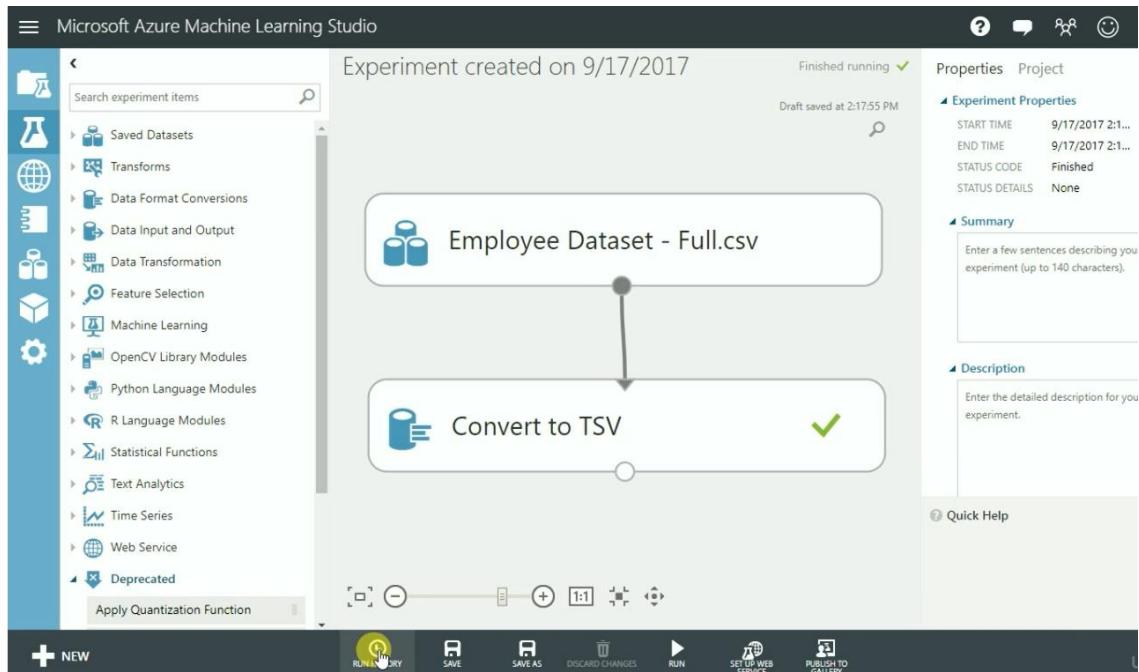
Can preview module as below using mini map



Can Perform actions like Zoom in, zoom out and make fit to screen click below



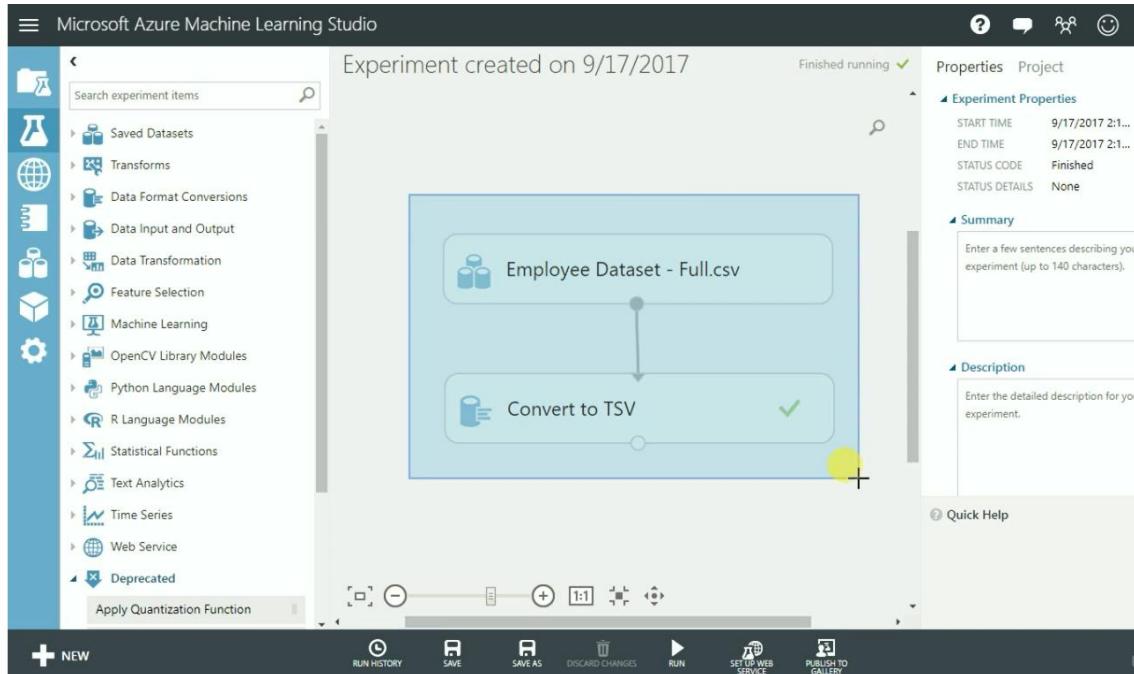
By clicking run history, can view experiment history



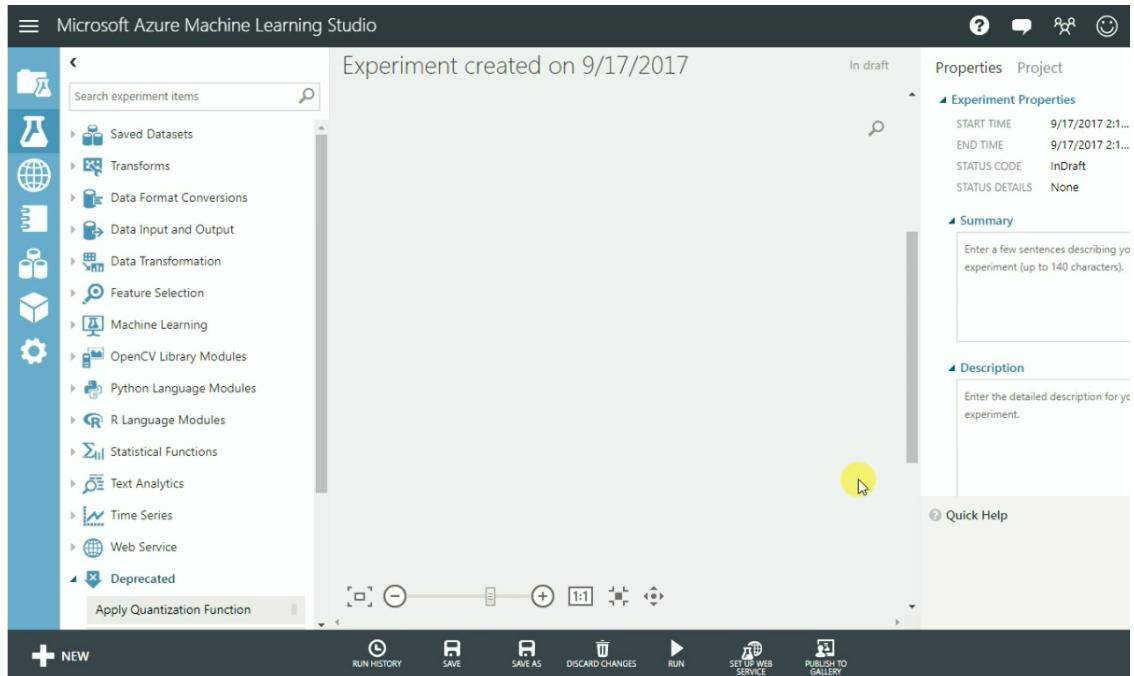
The screenshot shows the Microsoft Azure Machine Learning Studio interface with the title "experiment created on 9/17/2017". On the left, there's a sidebar with various icons. The main area displays a table titled "Run History" with the following data:

NAME	STATE	STATUS	START TIME	END TIME
Experiment created on 9/17/2017	Editable	Finished	9/17/2017 2:16:54 PM	9/17/2017 2:16:54 PM
Experiment created on 9/17/2017	Locked	Finished	9/17/2017 2:16:54 PM	9/17/2017 2:16:54 PM

To delete the modules simply select the modules and press delete for free space

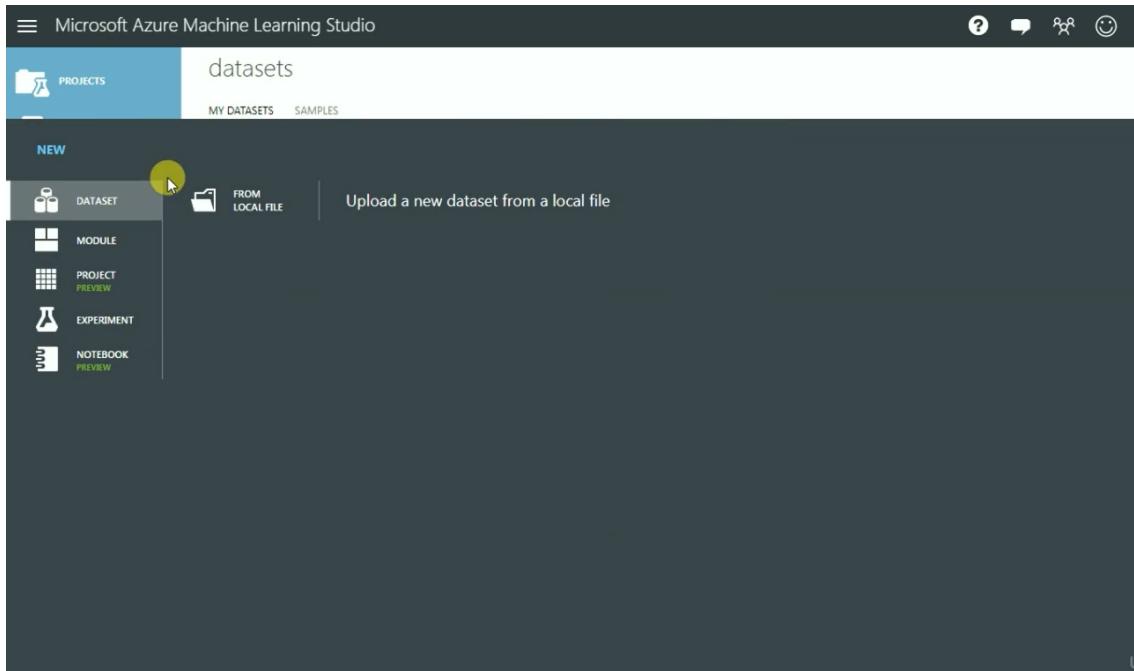


## After deletion

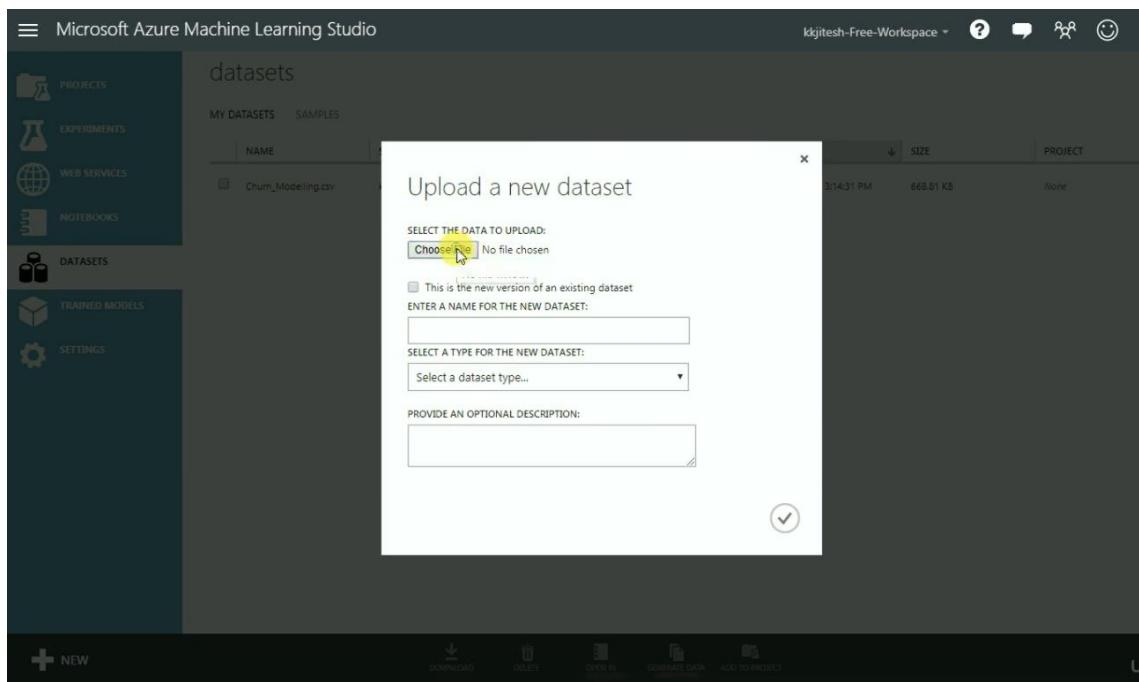


## Unpacking the Zipped Dataset

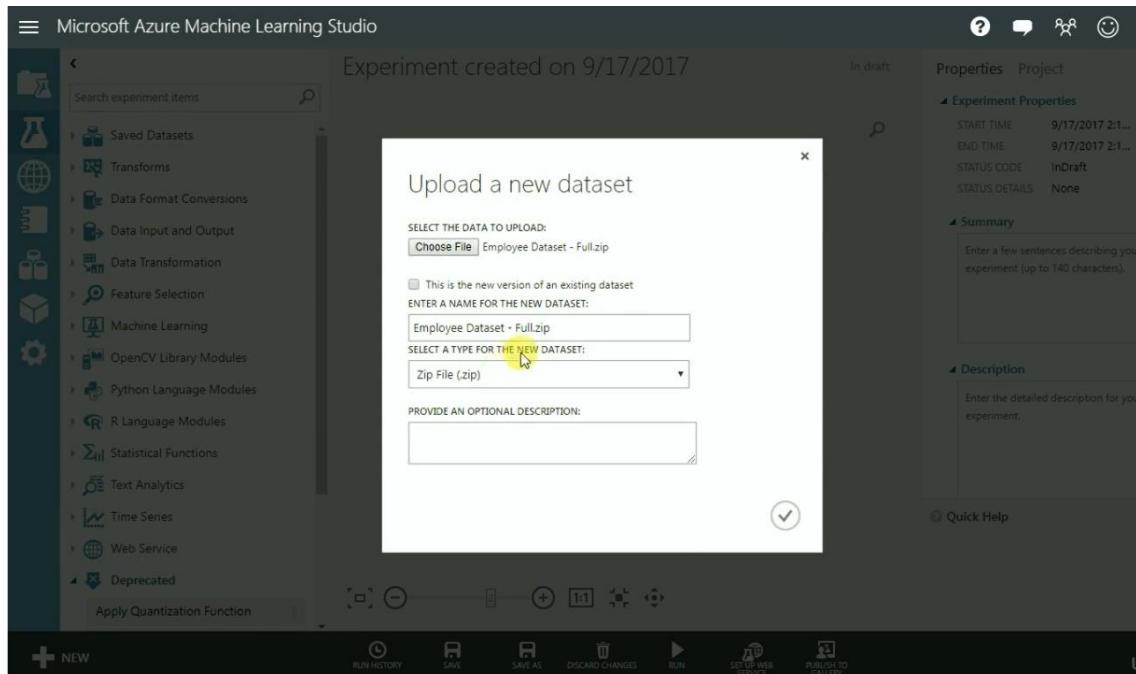
Go to dataset and click From Local file



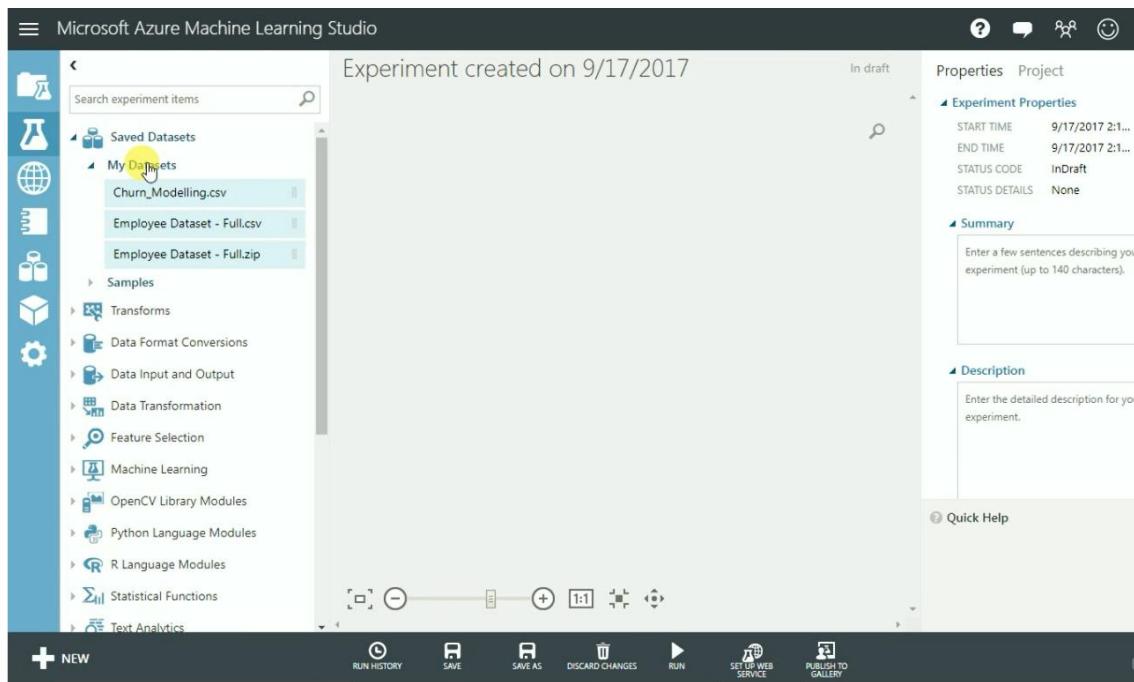
Choose zipped file from local drive



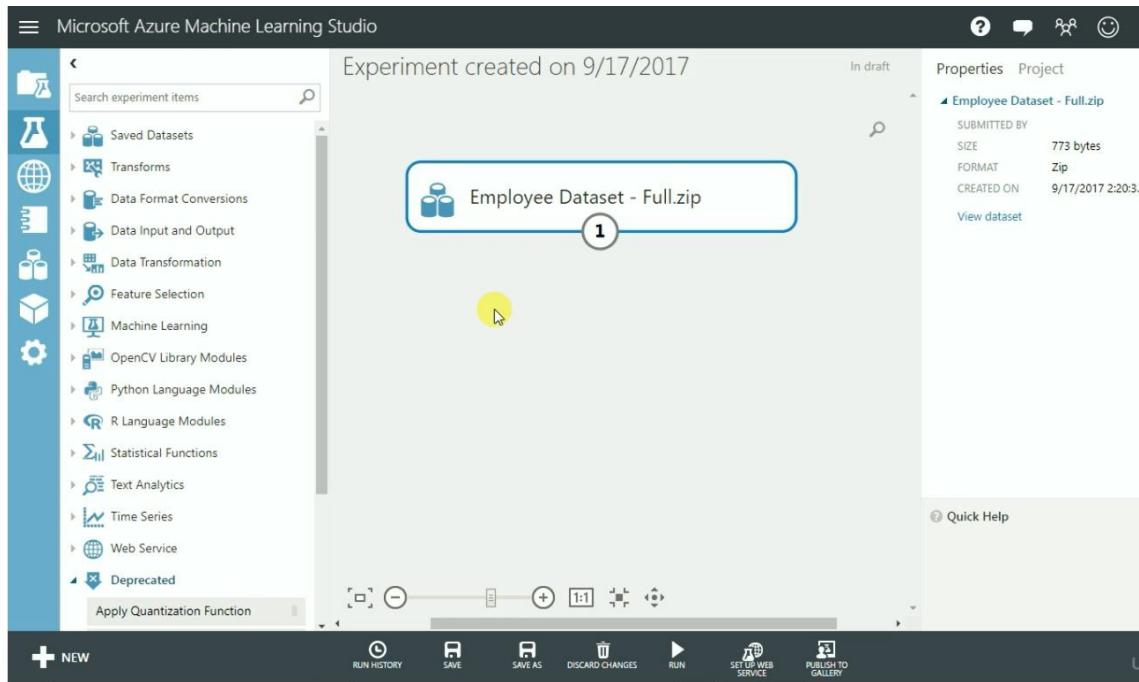
## Upload the zip file



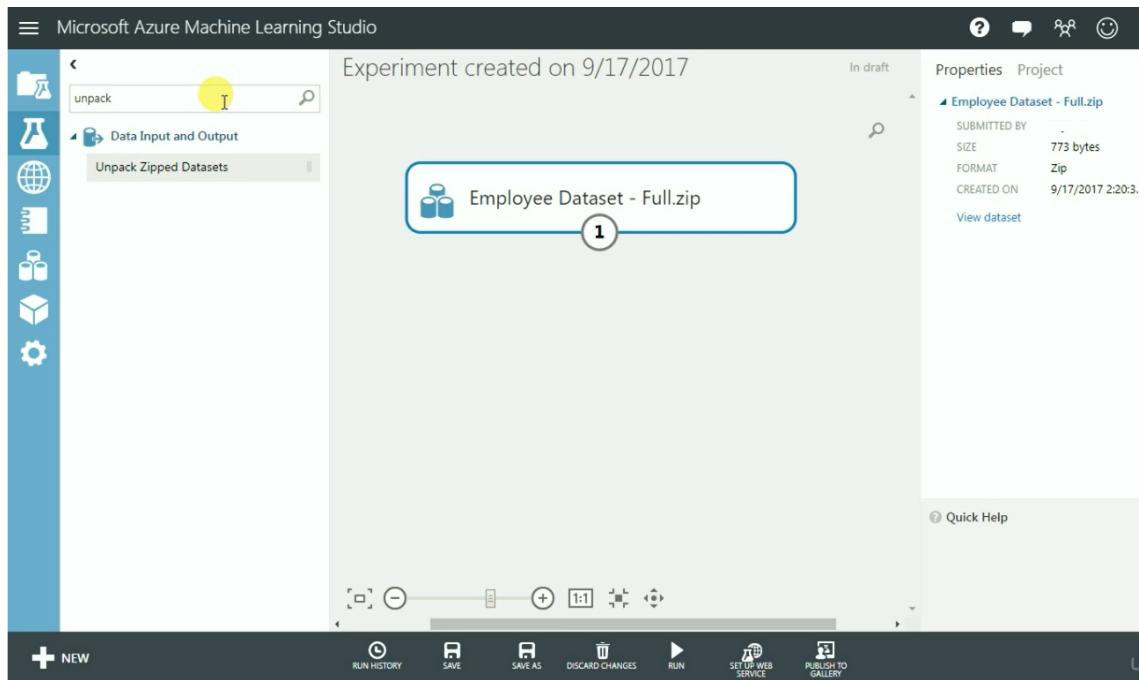
Click on saved datasets → My datasets → Zipped file



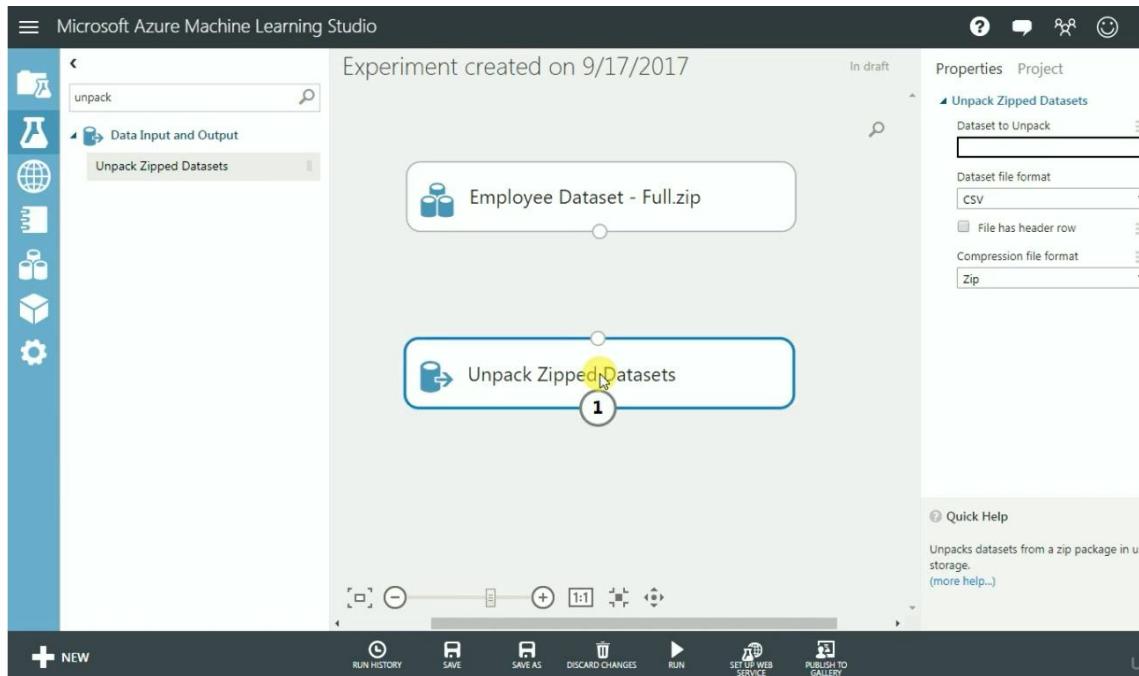
## Drag and drop the dataset



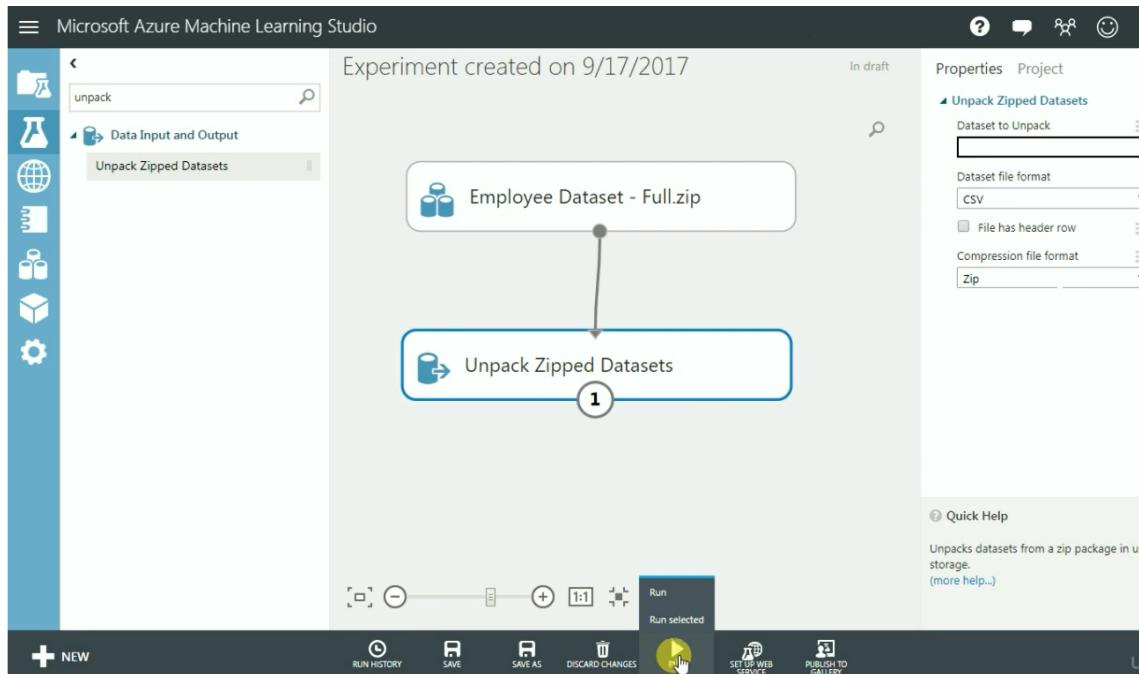
## Search for unpack Zipped datasets



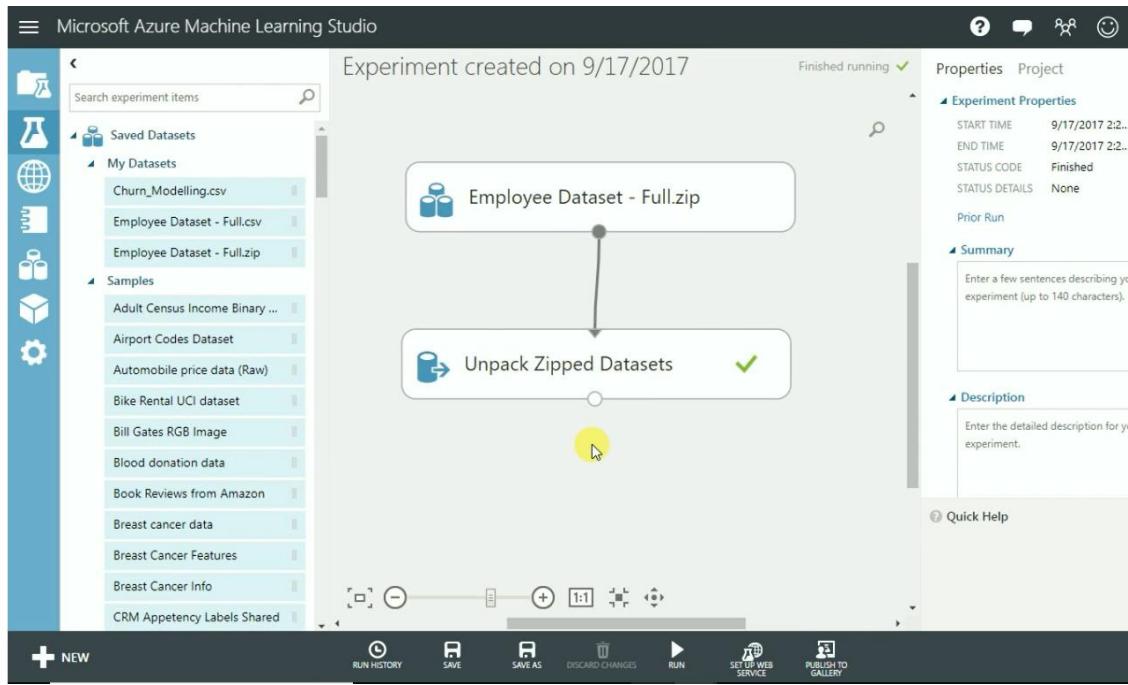
## Drag and drop Unzip dataset



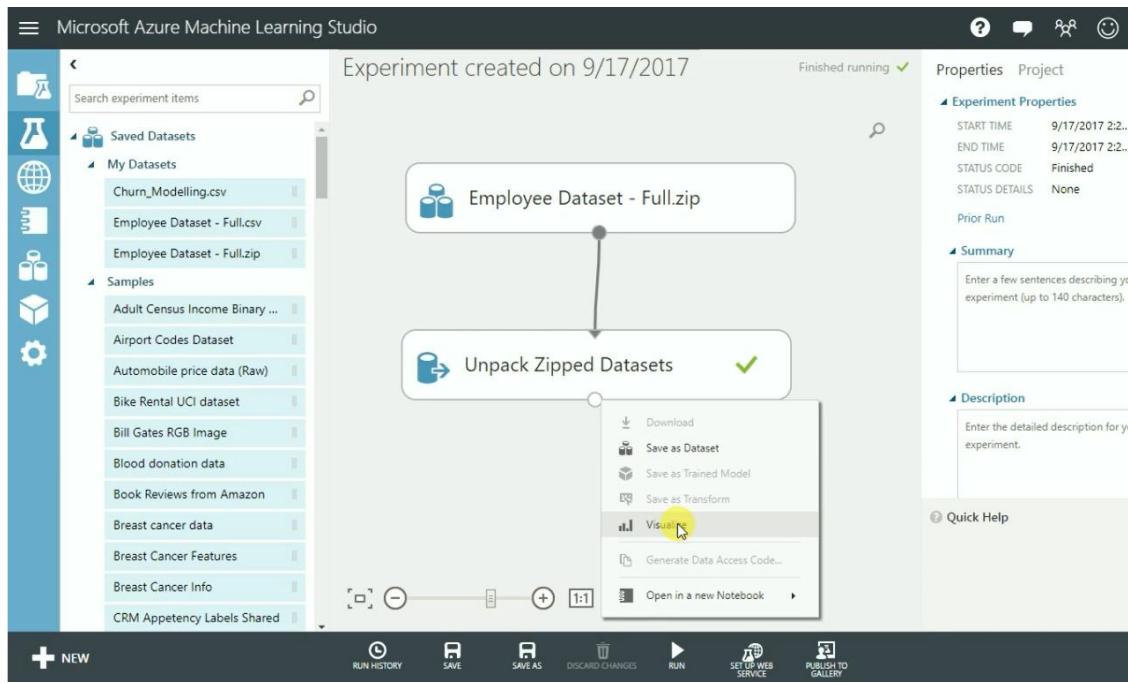
## Connect the output and input nodes and run the module



## Post successful execution



Right click on output node and visualize



Visualize and check the output. However, result obtained header shows col1, col2 etc. to sort out this refer next slide.

The screenshot shows the Microsoft Azure Machine Learning Studio interface. In the center, there is a table visualization of a dataset named "Results dataset1". The table has 26 rows and 11 columns, with headers "Col1" through "Col11". The first few rows of data are visible:

Employee Name	Age	Last Working Day	Department	Education	Gender	Ma
Jitesh	41	31-12-9999	Training	Masters	Male	Si
Sanjiti	49	31-12-9999	Sales	Masters	Male	Ma
John	37	31-12-9999	R&D	Doctorate	Male	Si
Sandra	33	31-12-9999	Software Development	Undergraduate	Female	Ma
Madhu	27	31-12-9999	R&D	Masters	Male	Ma
Robert	32	31-12-9999	R&D	Masters	Male	Si
Megan	59	31-12-9999	Software Development	Masters	Female	Ma

On the right side of the table, there are sections for "Statistics" and "Visualizations". The "Visualizations" section shows small bar charts for each column. A yellow circle highlights the scroll bar on the right side of the table area.

Click on file has a header row as shown in right side and run the module

The screenshot shows the Microsoft Azure Machine Learning Studio interface. On the left, the "Saved Datasets" section is expanded, showing various datasets like "Churn\_Modelling.csv", "Employee Dataset - Full.csv", and "Employee Dataset - Full.zip".

In the center, there is a workflow diagram. An arrow points from "Employee Dataset - Full.zip" to the "Unpack Zipped Datasets" module. This module is highlighted with a yellow circle and has a green checkmark icon. The status of the module is "1".

On the right, the "Properties" panel is open for the "Unpack Zipped Datasets" module. It shows the following settings:

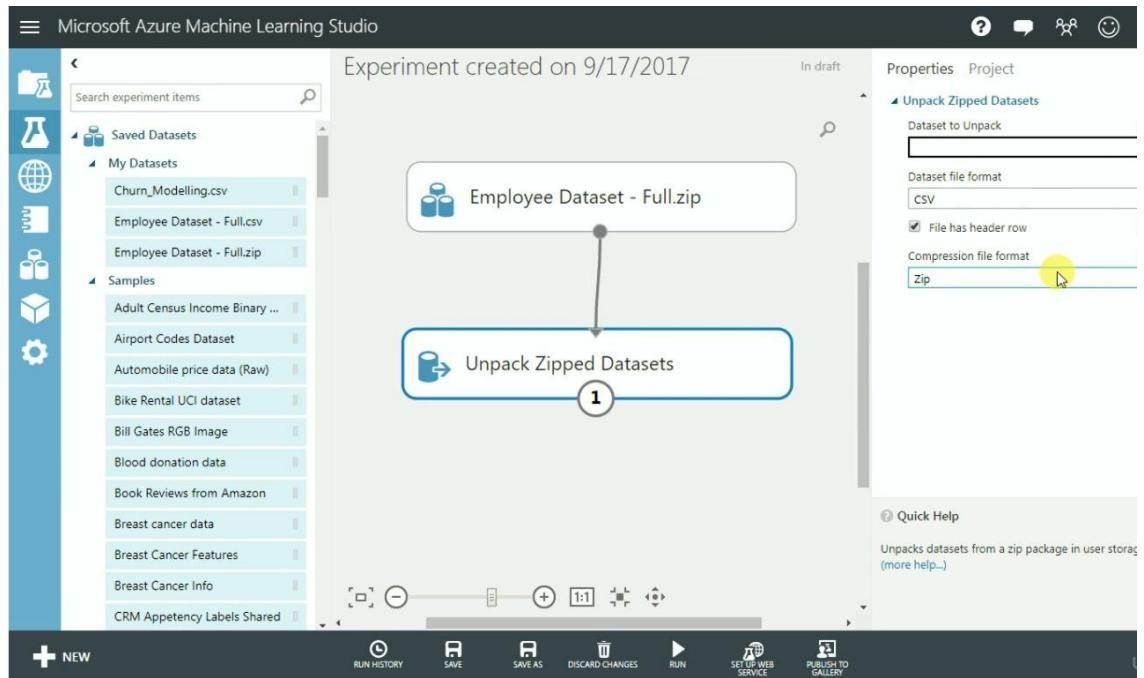
- Dataset to Unpack:** Employee Dataset - Full.zip
- Dataset file format:** CSV
- File has header row:**  (highlighted with a yellow circle)
- Compression file format:** Zip

Below these settings, the module's status is listed:

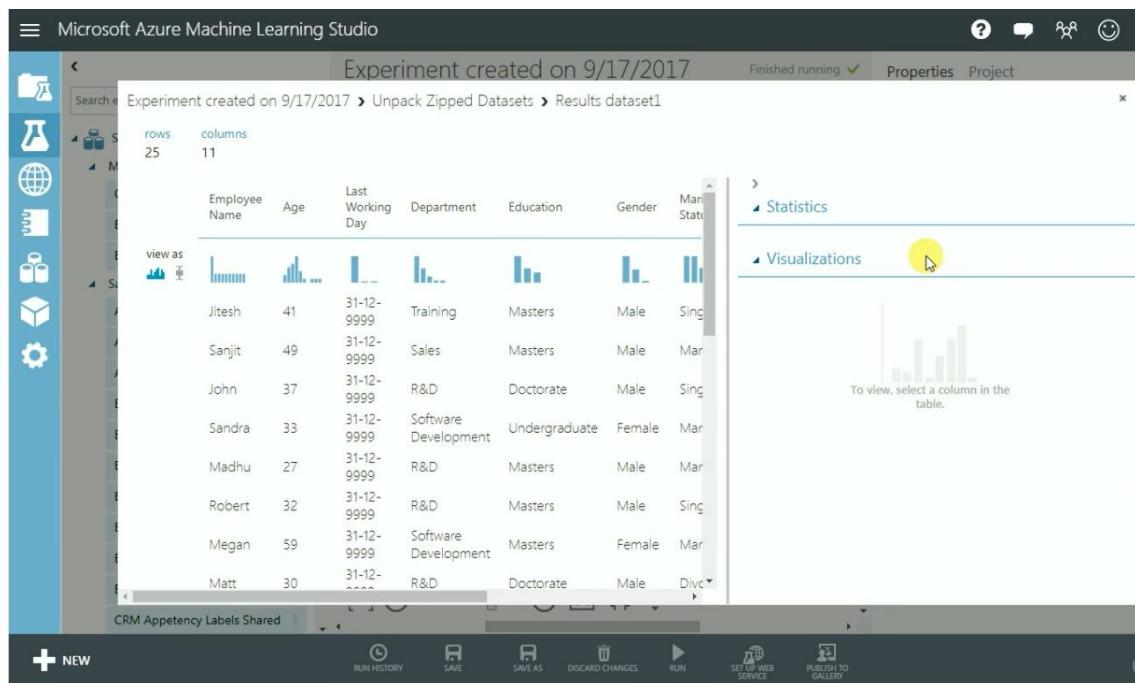
- START TIME: 9/17/2017 2:22:3
- END TIME: 9/17/2017 2:22:4
- ELAPSED TIME: 0:00:07.964
- STATUS CODE: Finished
- STATUS DETAILS: None

A "View output log" link is also present. At the bottom of the properties panel, there is a "Quick Help" section with the text: "Unpacks datasets from a zip package in u storage. (more help...)"

Mark the file has header row and run the module again

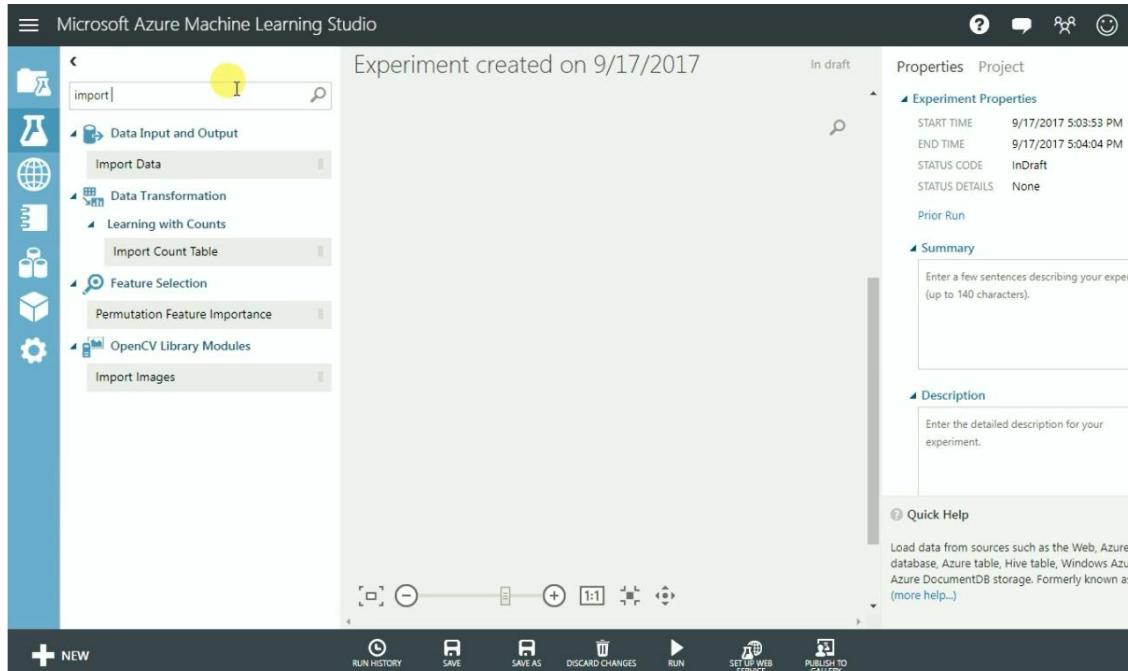


Post execution find the output obtained is correct

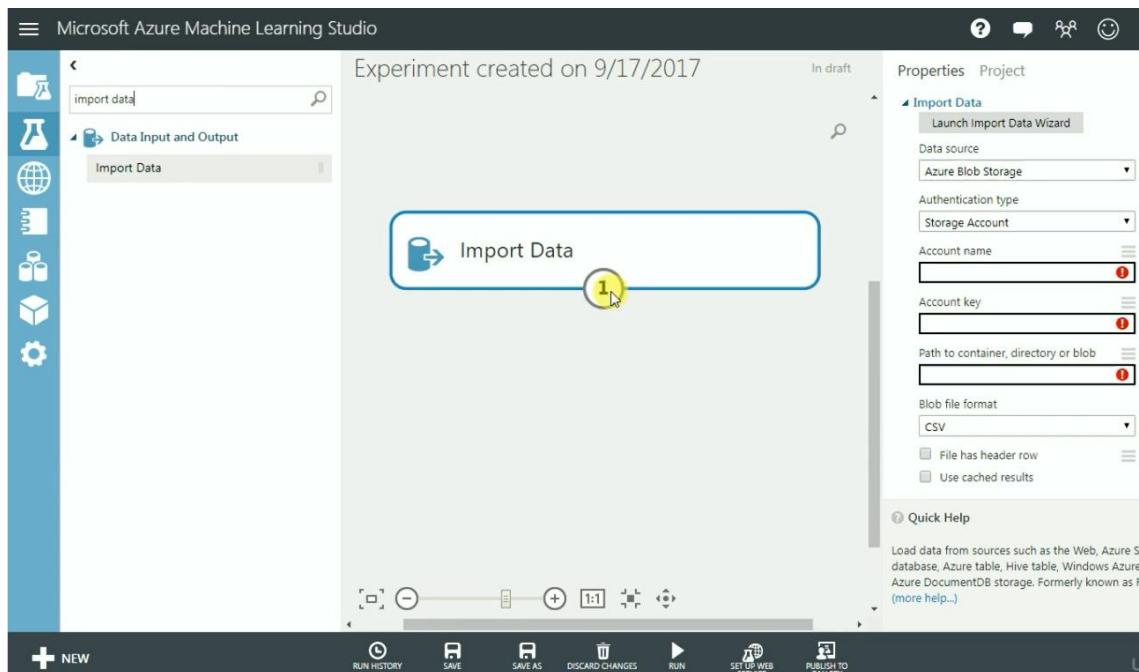


# Importing Data from Multiple Sources

Search for the import data



Drag and drop to the canvas



Can import data from various sources as shown below

Select Web URL via HTTP from the dropdown

The screenshot shows the Microsoft Azure Machine Learning Studio interface. On the left, there's a vertical toolbar with icons for file operations, data input/output, and other tools. The main workspace is titled "Experiment created on 9/17/2017" and has a status "In draft". In the center, there's a large button labeled "Import Data" with a blue arrow icon. To the right of the workspace, there's a "Properties" panel. Under the "Import Data" section, the "Data source" dropdown menu is open, showing options like "Azure Blob Storage", "Web URL via HTTP", "Hive Query", and "Azure SQL Database". The "Azure SQL Database" option is highlighted with a yellow circle. Below the dropdown, there are fields for "Path to container, directory or blob" and "Blob file format" set to "CSV". There are also checkboxes for "File has header row" and "Use cached results". A "Quick Help" section provides a brief description of the import functionality.

Input the URL from where data to be imported in next column

URL : <https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data>

This screenshot is similar to the one above, showing the "Import Data" interface in Microsoft Azure Machine Learning Studio. The "Data source" dropdown is now set to "Web URL via HTTP". The "Data source URL" field contains the value "https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data", which is highlighted with a yellow circle. The "Data format" dropdown is set to "CSV". The "Quick Help" section at the bottom remains the same, providing information about loading data from various sources.

## Data found in URL

```

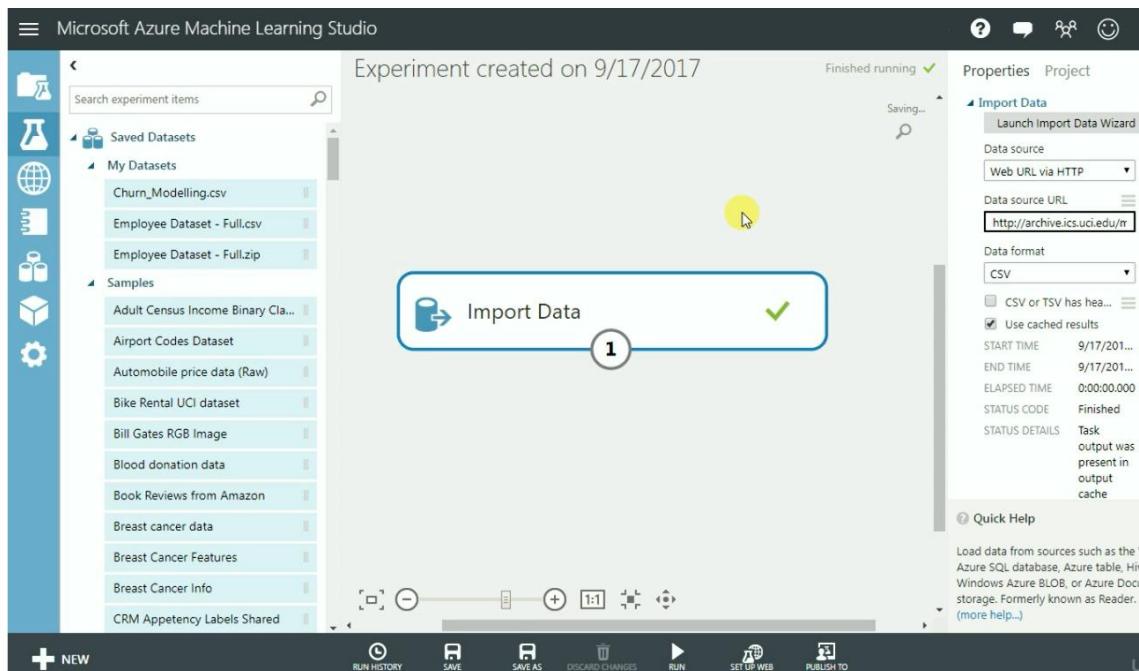
39, State-gov, 77516, Bachelors, 13, Never-married, Adm-clerical, Not-in-family, White, Male, 2174, 0, 40, United-States, <=50K
50, Self-emp-not-inc, 83311, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 13, United-States, <=50K
50, Private, 215646, HS-grad, 9, Divorced, Handlers-cleaners, Not-in-family, White, Male, 0, 0, 40, United-States, <=50K
38, Private, 234721, 11th, 7, Married-civ-spouse, Handlers-cleaners, Husband, Black, Male, 0, 0, 40, United-States, <=50K
53, Private, 338409, Bachelors, 13, Married-civ-spouse, Prof-specialty, Wife, Black, Female, 0, 0, 40, Cuba, <=50K
28, Private, 284582, Masters, 14, Married-civ-spouse, Exec-managerial, Wife, White, Female, 0, 0, 40, United-States, <=50K
37, Private, 160187, 9th, 5, Married-spouse-absent, Other-service, Not-in-family, Black, Female, 0, 0, 16, Jamaica, <=50K
49, Private, 45781, Masters, 14, Never-married, Prof-specialty, Not-in-family, White, Female, 14084, 0, 50, United-States, <=50K
31, Private, 159449, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 5178, 0, 40, United-States, <=50K
42, Private, 280464, Some-college, 10, Married-civ-spouse, Exec-managerial, Husband, Black, Male, 0, 0, 80, United-States, <=50K
37, Private, 140997, Some-college, 10, Married-civ-spouse, Exec-managerial, Husband, Asian-Pac-Islander, Male, 0, 0, 40, India, >50K
36, State-gov, 205019, Bachelors, 13, Never-married, Adm-clerical, Own-child, White, Female, 0, 0, 30, United-States, <=50K
23, Private, 12172, Bachelors, 13, Never-married, Sales, Not-in-family, Black, Male, 0, 0, 50, United-States, <=50K
32, Private, 205019, Bachelors, 13, Never-married, Sales, Own-child, White, Male, 0, 0, 40, ?, >50K
40, Private, 12172, Assoc-voc, 13, Married-civ-spouse, Craft-repair, Husband, Asian-Pac-Islander, Male, 0, 0, 40, ?, >50K
34, Private, 245487, 7th-8th, 4, Married-civ-spouse, Transport-moving, Husband, Amer-Indian-Eskimo, Male, 0, 0, 45, Mexico, <=50K
25, Self-emp-not-inc, 209642, HS-grad, 9, Never-married, Farming-fishing, Own-child, White, Male, 0, 0, 35, United-States, <=50K
32, Private, 168624, HS-grad, 9, Never-married, Machine-op-inspct, Unmarried, White, Male, 0, 0, 40, United-States, <=50K
38, Private, 28887, 11th, 7, Married-civ-spouse, Sales, Husband, White, Male, 0, 0, 50, United-States, <=50K
43, Self-emp-not-inc, 292175, Masters, 14, Divorced, Exec-managerial, Unmarried, White, Female, 0, 0, 45, United-States, >50K
40, Private, 193524, Doctorate, 16, Married-civ-spouse, Prof-specialty, Husband, White, Male, 0, 0, 60, United-States, >50K
54, Private, 302146, HS-grad, 9, Separated, Other-service, Unmarried, Black, Female, 0, 0, 20, United-States, <=50K
35, Federal-gov, 76845, 9th, 5, Married-civ-spouse, Farming-fishing, Husband, Black, Male, 0, 0, 40, United-States, <=50K
43, Private, 117037, 11th, 7, Married-civ-spouse, Transport-moving, Husband, White, Male, 0, 0, 2042, 40, United-States, <=50K
59, Private, 109015, HS-grad, 9, Divorced, Tech-support, Unmarried, White, Female, 0, 0, 40, United-States, <=50K
56, Local-gov, 216851, Bachelors, 13, Married-civ-spouse, Tech-support, Husband, White, Male, 0, 0, 40, United-States, >50K
19, Private, 168294, HS-grad, 9, Never-married, Craft-repair, Own-child, White, Male, 0, 0, 40, United-States, <=50K
54, ?, 180211, Some-college, 10, Married-civ-spouse, ?, Husband, Asian-Pac-Islander, Male, 0, 0, 60, South, <50K
39, Private, 367260, HS-grad, 9, Divorced, Exec-managerial, Not-in-family, White, Male, 0, 0, 80, United-States, <=50K
49, Private, 193524, HS-grad, 9, Married-civ-spouse, Craft-repair, Husband, White, Male, 0, 0, 40, United-States, <=50K
23, Local-gov, 100794, Assoc-acdm, 12, Never-married, Protective-service, Not-in-family, White, Male, 0, 0, 52, United-States, <=50K
20, Private, 266015, Some-college, 10, Never-married, Sales, Own-child, Black, Male, 0, 0, 40, United-States, <=50K
45, Private, 386940, Bachelors, 13, Divorced, Exec-managerial, Own-child, White, Male, 0, 0, 1408, 40, United-States, <=50K
30, Federal-gov, 590951, Some-college, 10, Married-civ-spouse, Adm-clerical, Own-child, White, Male, 0, 0, 40, United-States, <=50K
22, State-gov, 311512, Some-college, 10, Married-civ-spouse, Other-service, Husband, Black, Male, 0, 0, 15, United-States, <=50K
48, Private, 242406, 11th, 7, Never-married, Machine-op-inspct, Unmarried, White, Male, 0, 0, 40, Puerto-Rico, <=50K
21, Private, 197200, Some-college, 10, Never-married, Machine-op-inspct, Own-child, White, Male, 0, 0, 40, United-States, <=50K
19, Private, 544091, HS-grad, 9, Married-af-spouse, Adm-clerical, Wife, White, Female, 0, 0, 25, United-States, <=50K
31, Private, 84154, Some-college, 10, Married-civ-spouse, Sales, Husband, White, Male, 0, 0, 38, ?, >50K
48, Self-emp-not-inc, 265477, Assoc-acdm, 12, Married-civ-spouse, Prof-specialty, Husband, White, Male, 0, 0, 40, United-States, <=50K
31, Private, 507875, 9th, 5, Married-civ-spouse, Machine-op-inspct, Husband, White, Male, 0, 0, 43, United-States, <=50K
53, Self-emp-not-inc, 88506, Bachelors, 13, Married-civ-spouse, Prof-specialty, Husband, White, Male, 0, 0, 40, United-States, <=50K
24, Private, 172987, Bachelors, 13, Married-civ-spouse, Tech-support, Husband, White, Male, 0, 0, 50, United-States, <=50K
49, Private, 94638, HS-grad, 9, Separated, Adm-clerical, Unmarried, White, Female, 0, 0, 40, United-States, <=50K
25, Private, 289980, HS-grad, 9, Never-married, Handlers-cleaners, Not-in-family, White, Male, 0, 0, 35, United-States, <=50K
57, Federal-gov, 337895, Bachelors, 13, Married-civ-spouse, Prof-specialty, Husband, Black, Male, 0, 0, 40, United-States, >50K

```

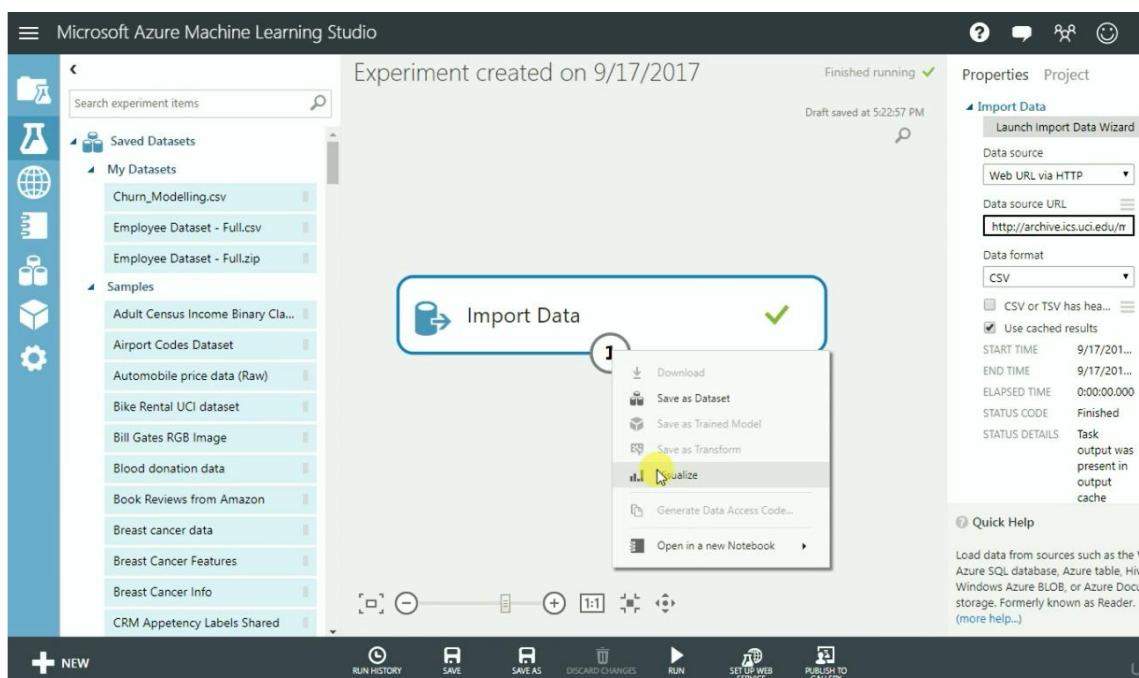
## Run the module for result

The screenshot shows the Microsoft Azure Machine Learning Studio interface. A central workspace contains a 'Import Data' module. A context menu is open over this module, with the 'Run selected' option highlighted. The background shows the studio's navigation bar, a properties panel on the right, and various other modules and toolbars.

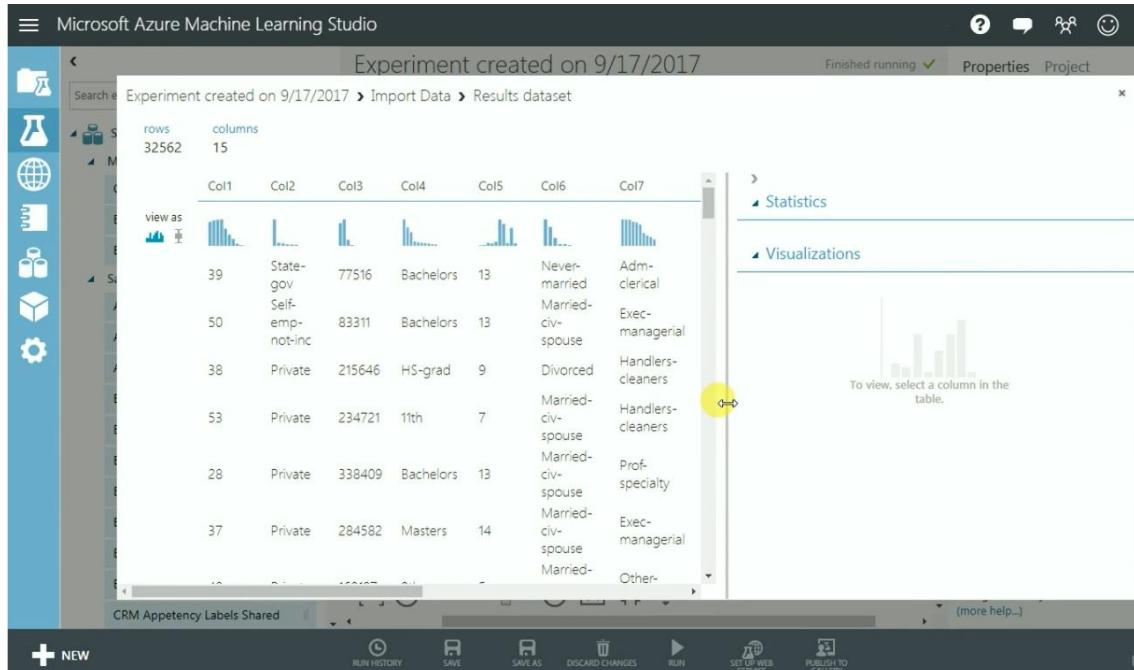
After successful execution visualize the data



Right click the node and visualize the data

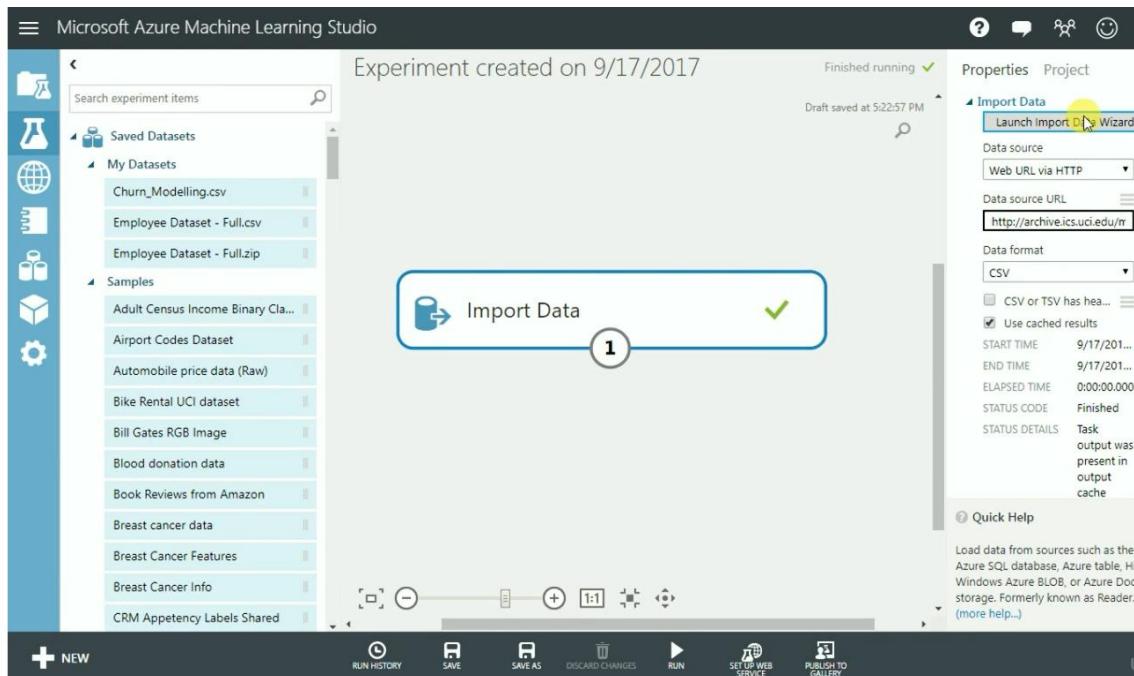


## Data imported from URL

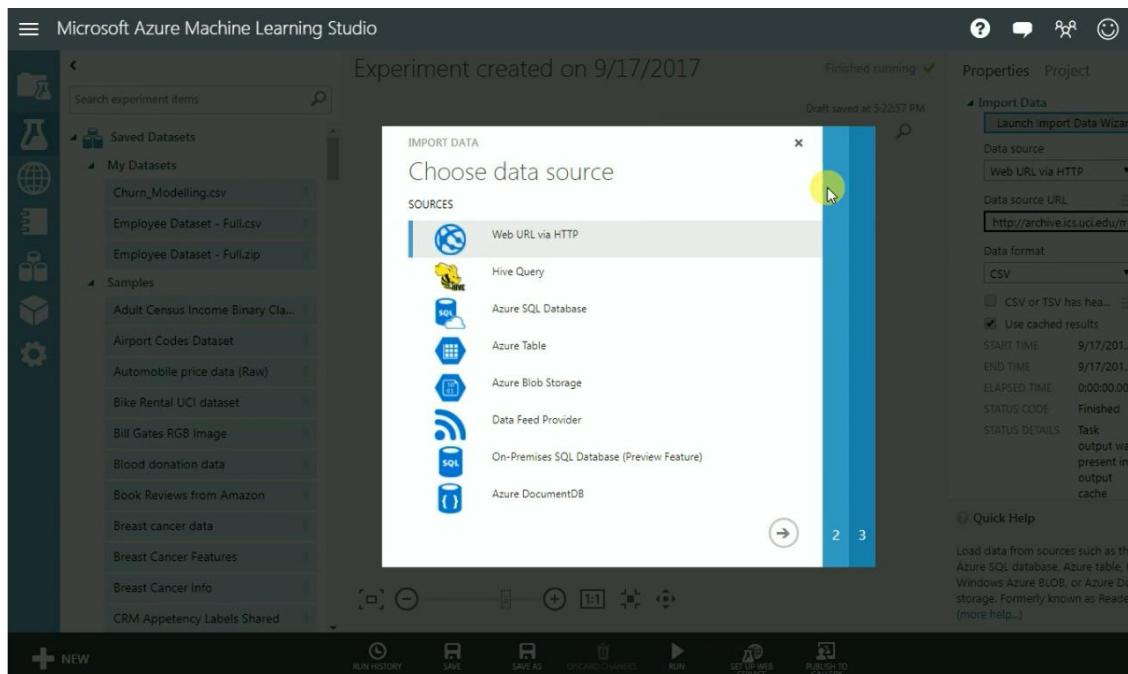


## Launch import data wizard

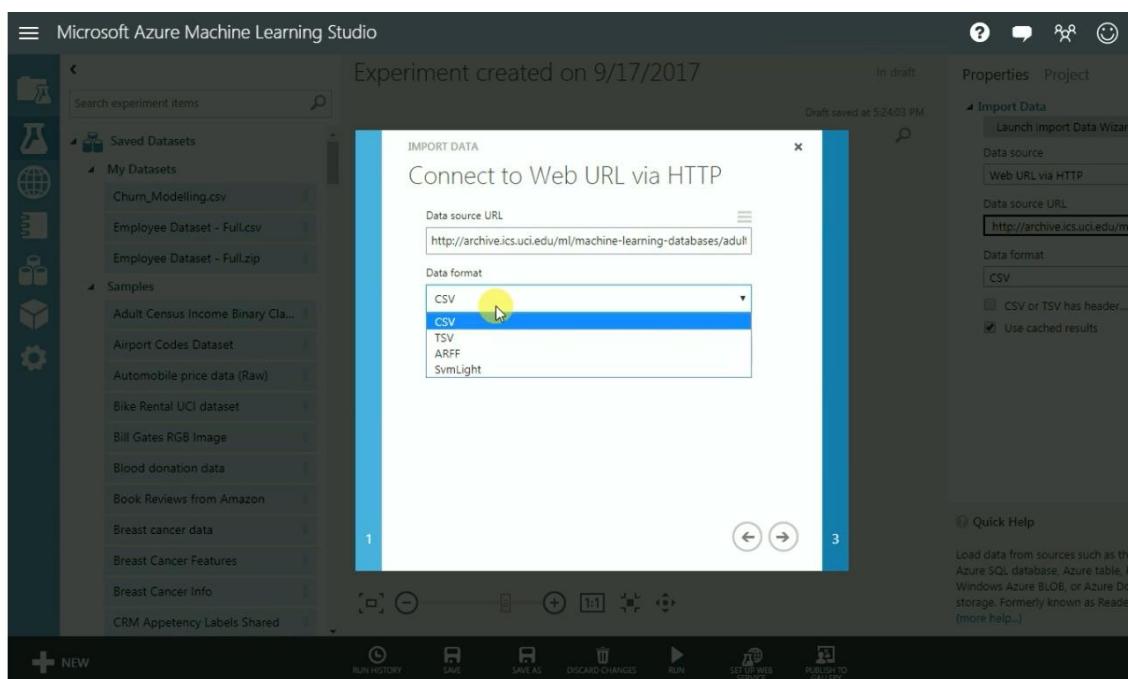
Click Launch import data wizard in the right side



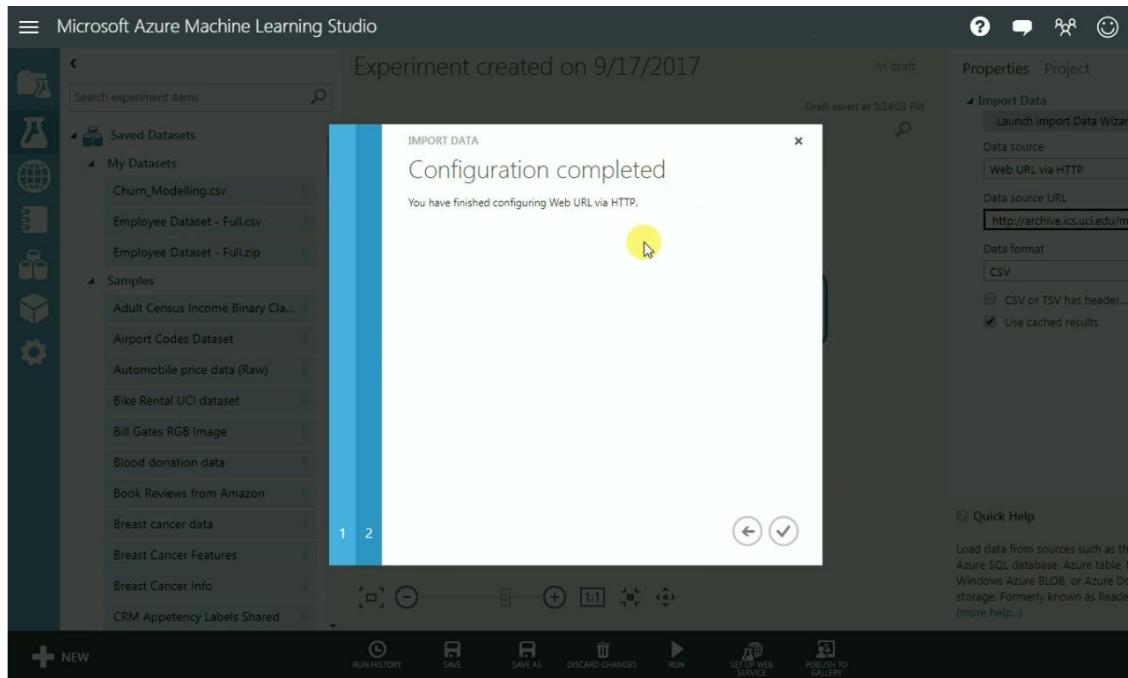
Select web URL from HTTP from the list and click on forward arrow



Enter the data source URL and data format

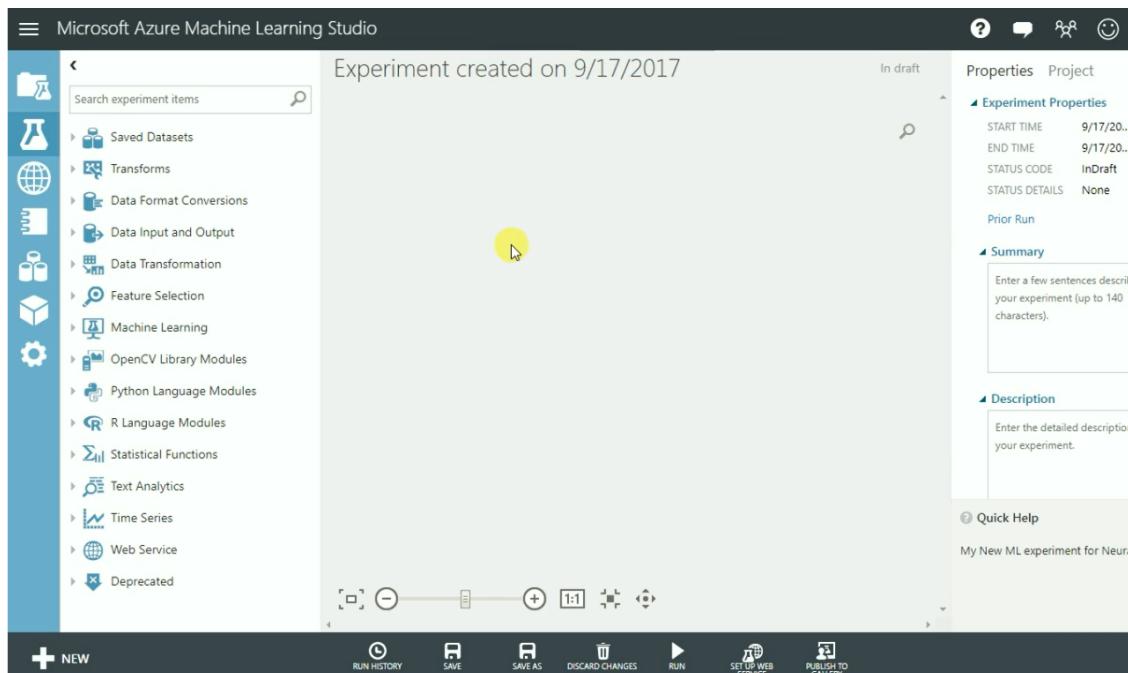


After successful configuration, you can run the module and get same result

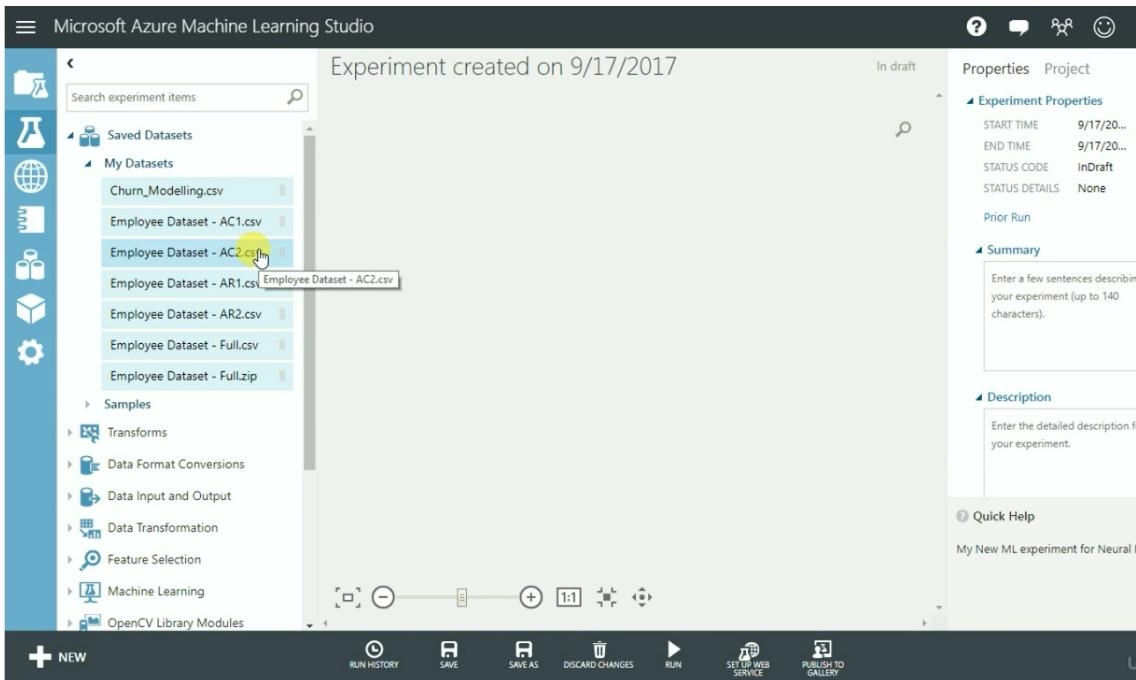


## Data Manipulation Using Add Columns Component

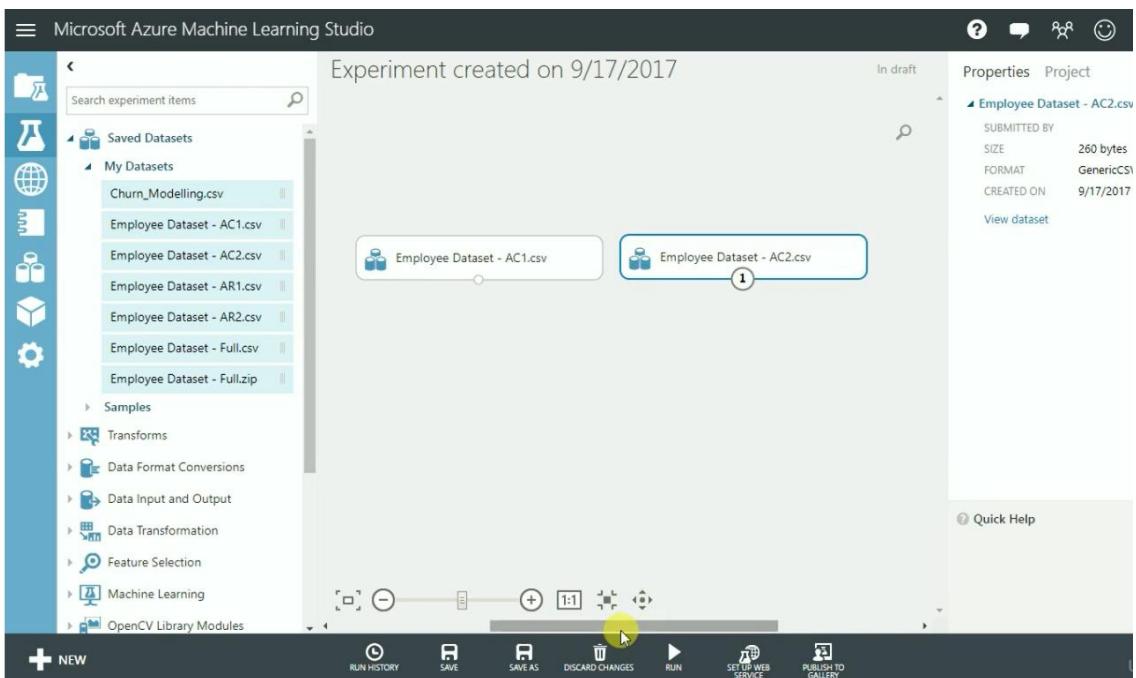
### Add Columns



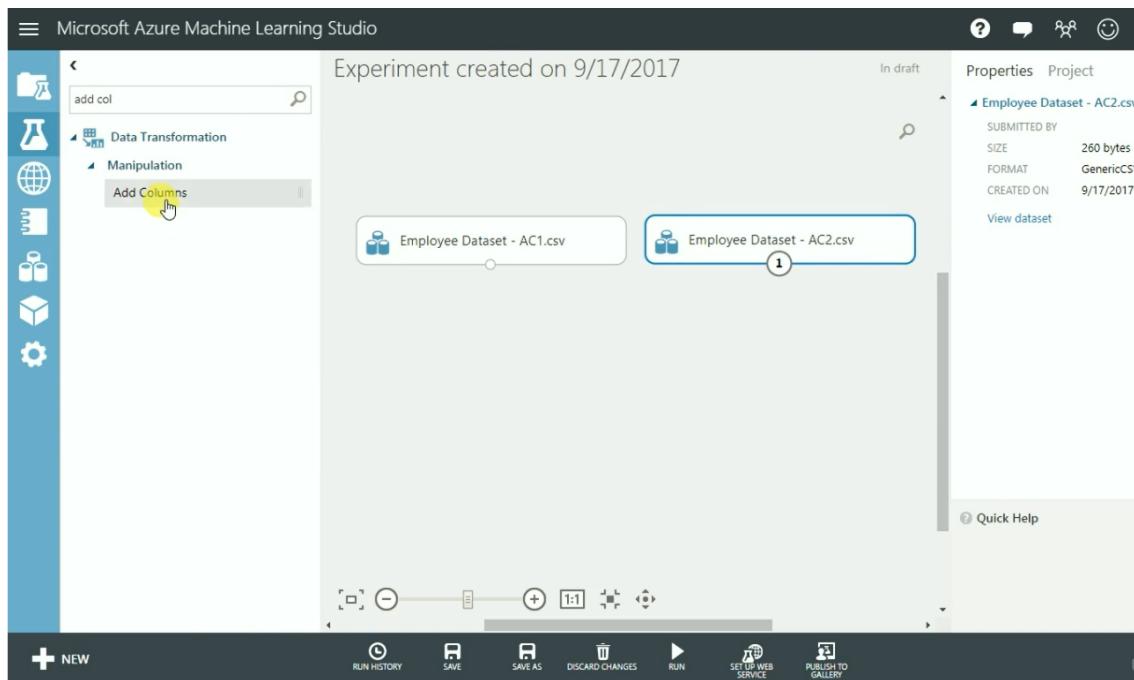
Select any two datasets



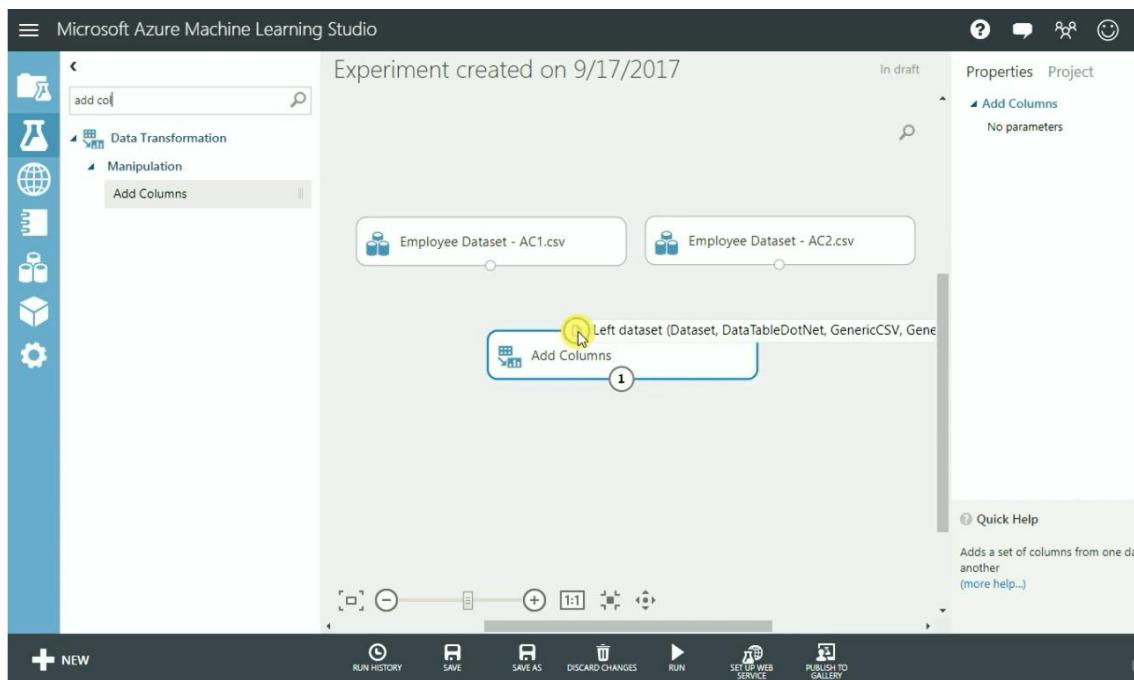
Drag and drop two datasets from the list



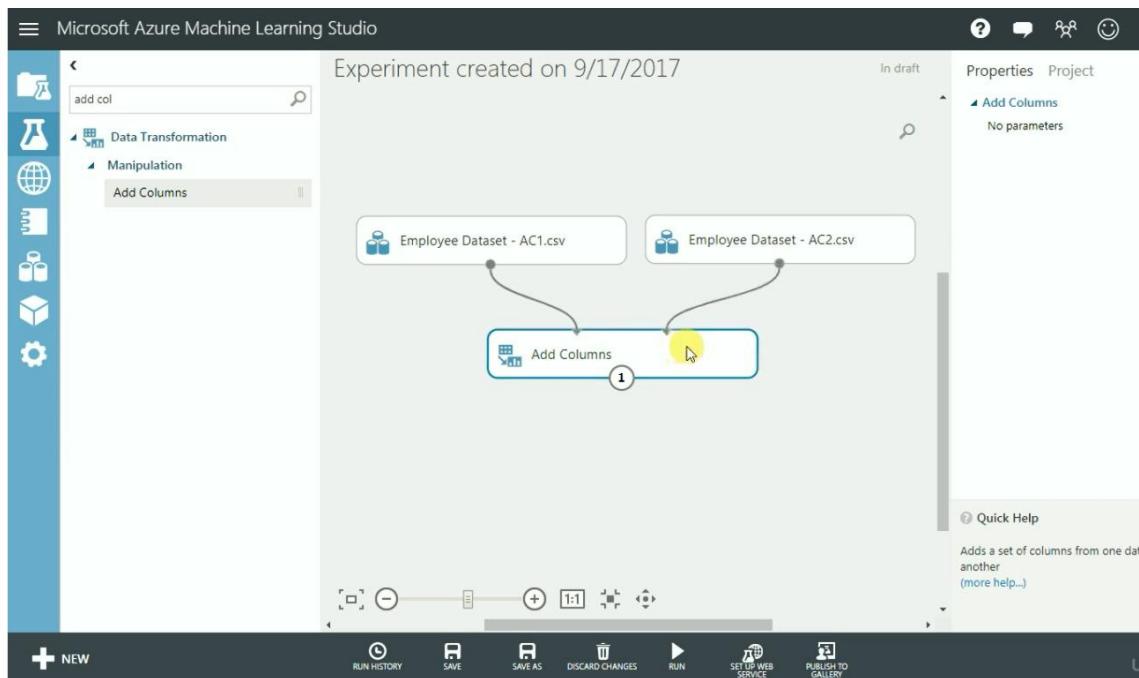
Search for add columns and drop the same in canvas



Add column has left dataset and Right dataset



Connect both csv files to Add Column Left and right input node



Visualize both csv files before execution to know its columns

Experiment created on 9/17/2017

rows 25 columns 8

Employee Name, Age, Last Working Day, Department, Education, Gender, Marital Status, Manager

view as

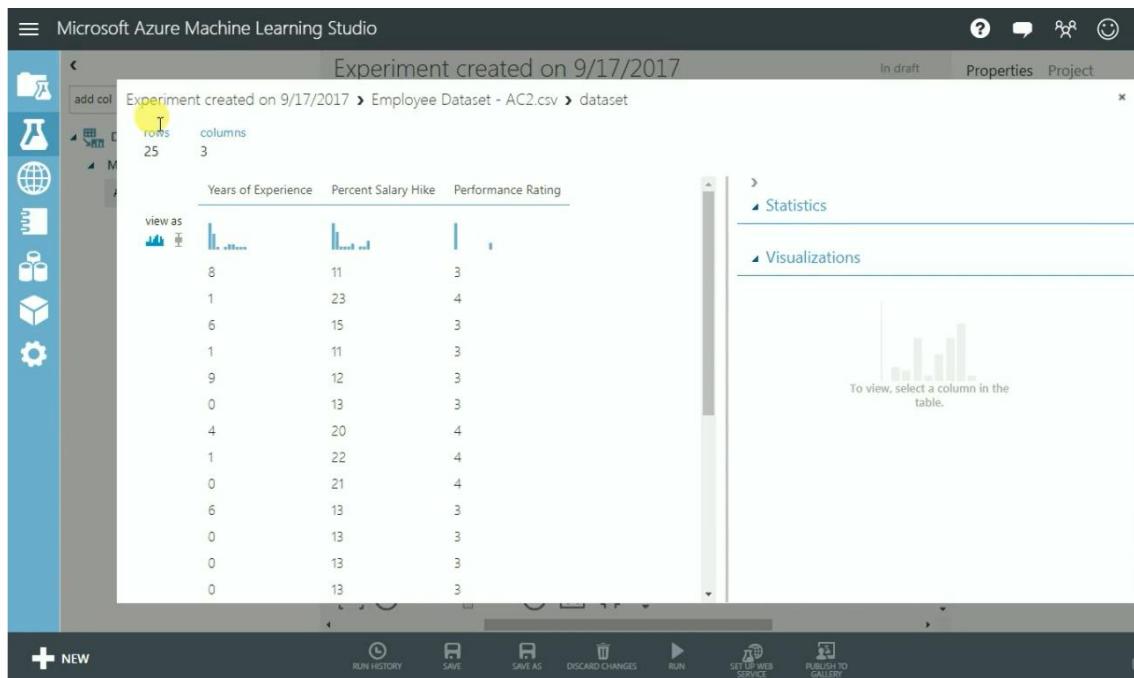
Employee Name	Age	Last Working Day	Department	Education	Gender	Marital Status	Manager
Jitesh	41	31-12-9999	Training	Masters	Male	Sing	
Sanjita	49	31-12-9999	Sales	Masters	Male	Mar	
John	37	31-12-9999	R&D	Doctorate	Male	Sing	
Sandra	33	31-12-9999	Software Development	Undergraduate	Female	Mar	
Madhu	27	31-12-9999	R&D	Masters	Male	Mar	
Robert	32	31-12-9999	R&D	Masters	Male	Sing	
Megan	59	31-12-9999	Software Development	Masters	Female	Mar	
Matt	30	31-12-2000	R&D	Doctorate	Male	Diva	

Statistics

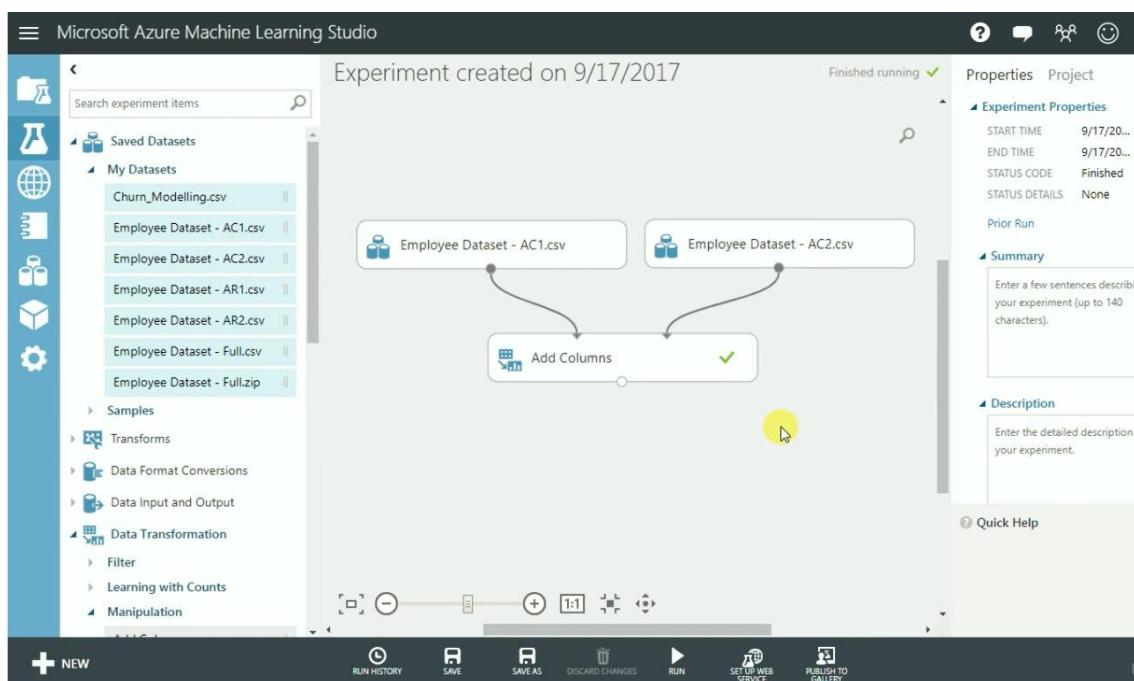
Visualizations

To view, select a column in the table.

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



Run the module and click visualize



Result is as expected that columns has been updated

Experiment created on 9/17/2017

rows 25 columns 11

Employee Name	Age	Last Working Day	Department	Education	Gender	Marital Status
Jitesh	41	31-12-9999	Training	Masters	Male	Sing
Sanjit	49	31-12-9999	Sales	Masters	Male	Mar
John	37	31-12-9999	R&D	Doctorate	Male	Sing
Sandra	33	31-12-9999	Software Development	Undergraduate	Female	Mar
Madhu	27	31-12-9999	R&D	Masters	Male	Mar
Robert	32	31-12-9999	R&D	Masters	Male	Sing
Megan	59	31-12-9999	Software Development	Masters	Female	Mar
Matt	30	31-12-2000	R&D	Doctorate	Male	Divc

Statistics

Visualizations

To view, select a column in the table.

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

## Data Manipulation Using Add Rows Component

### Add Rows

#### Select datasets for Adding rows

Experiment created on 9/17/2017

In draft

Properties Project

Experiment Properties

START TIME 9/17/2017  
END TIME 9/17/2017  
STATUS CODE InDraft  
STATUS DETAILS None

Prior Run

Summary

Description

Quick Help

Churn\_Modelling.csv

Employee Dataset - AC1.csv

Employee Dataset - AC2.csv

Employee Dataset - AR1.csv

Employee Dataset - AR2.csv

Employee Dataset - Full.csv

Employee Dataset - Full.zip

Filter

Learning with Counts

Manipulation

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

Drag and drop selected datasets

Microsoft Azure Machine Learning Studio

Experiment created on 9/17/2017

In draft

Properties Project

Experiment Properties

- START TIME 9/17/20...
- END TIME 9/17/20...
- STATUS CODE InDraft
- STATUS DETAILS None

Prior Run

Summary

Description

Enter a few sentences describing your experiment (up to 140 characters).

Enter the detailed description for your experiment.

Quick Help

Search experiment items

Saved Datasets

My Datasets

- Churn\_Modelling.csv
- Employee Dataset - AC1.csv
- Employee Dataset - AC2.csv
- Employee Dataset - AR1.csv
- Employee Dataset - AR2.csv
- Employee Dataset - Full.csv
- Employee Dataset - Full.zip

Samples

Transforms

Data Format Conversions

Data Input and Output

Data Transformation

- Filter
- Learning with Counts
- Manipulation

Add Step

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

The screenshot shows the Microsoft Azure Machine Learning Studio interface. On the left, there's a sidebar with various icons for saved datasets, samples, transforms, data format conversions, data input/output, and data transformation steps. The main workspace displays two datasets: 'Employee Dataset - AR1.csv' and 'Employee Dataset - AR2.csv'. A yellow circle highlights the connection between these two datasets. The top right shows experiment properties like start and end times, status code (InDraft), and status details (None). The bottom right has buttons for running history, saving, discarding changes, running, setting up a web service, and publishing to a gallery.

Visualize both csv before executing

Microsoft Azure Machine Learning Studio

Experiment created on 9/17/2017

In draft Properties Project

Experiment created on 9/17/2017 > Employee Dataset - AR1.csv > dataset

rows columns

8 11

	Employee Name	Age	Last Working Day	Department	Education	Gender	Marital Status
New as	Jitesh	41	31-12-9999	Training	Masters	Male	Single
View as	Sanjit	49	31-12-9999	Sales	Masters	Male	Married
Save as	John	37	31-12-9999	R&D	Doctorate	Male	Single
Transform	Sandra	33	31-12-9999	Software Development	Undergraduate	Female	Married
Data	Madhu	27	31-12-9999	R&D	Masters	Male	Married
Manipulation	Robert	32	31-12-9999	R&D	Masters	Male	Single
Filter	Megan	59	31-12-9999	Software Development	Masters	Female	Married
Load	Matt	30	31-12-9999	R&D	Doctorate	Male	Divorced

Statistics

Visualizations

To view, select a column in the table.

Run History Save As Discard Changes Run Set Up Web Service Publish To Gallery

The screenshot shows the Microsoft Azure Machine Learning Studio interface. The main workspace displays a dataset named 'Employee Dataset - AR1.csv'. The table shows 8 rows and 11 columns with columns for Employee Name, Age, Last Working Day, Department, Education, Gender, and Marital Status. A yellow circle highlights the 'Run History' button at the bottom. The right side of the screen shows 'Statistics' and 'Visualizations' sections, with a note 'To view, select a column in the table.' The top right shows experiment properties and project settings.

Microsoft Azure Machine Learning Studio

Experiment created on 9/17/2017

In draft Properties Project

Experiment created on 9/17/2017 > Employee Dataset - AR2.csv > dataset

rows 17 columns 11

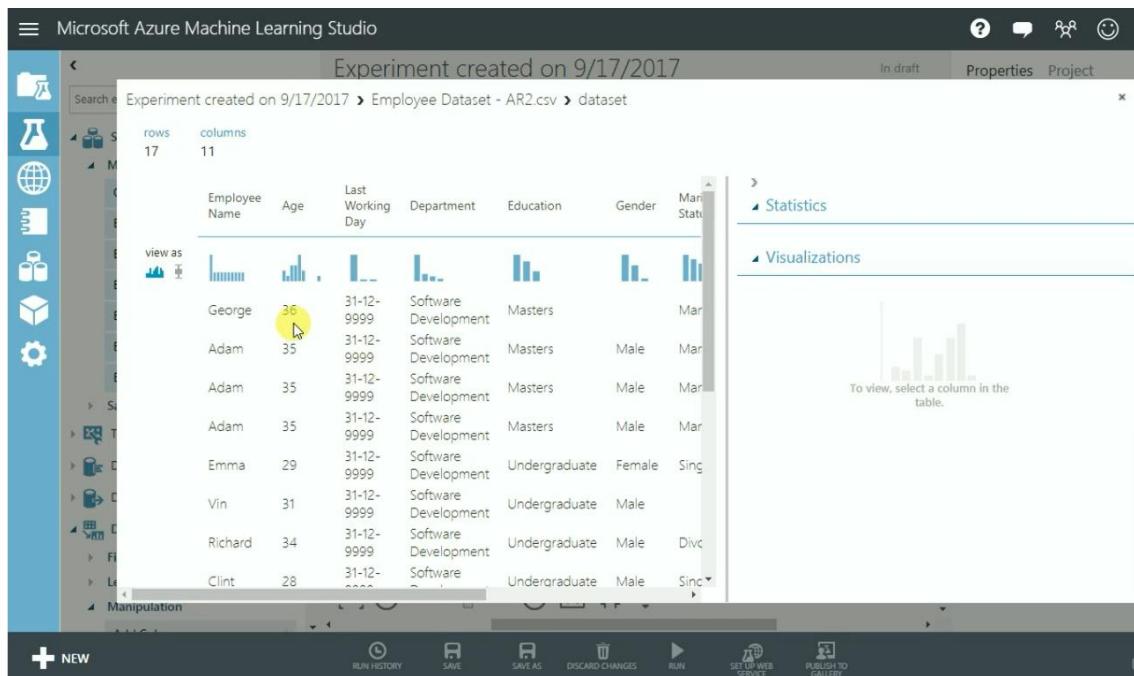
Employee Name Age Last Working Day Department Education Gender Mar Stat

Employee Name	Age	Last Working Day	Department	Education	Gender	Mar Stat
George	36	31-12-9999	Software Development	Masters	Male	Married
Adam	35	31-12-9999	Software Development	Masters	Male	Married
Adam	35	31-12-9999	Software Development	Masters	Male	Married
Adam	35	31-12-9999	Software Development	Masters	Male	Married
Emma	29	31-12-9999	Software Development	Undergraduate	Female	Single
Vin	31	31-12-9999	Software Development	Undergraduate	Male	
Richard	34	31-12-9999	Software Development	Undergraduate	Male	Divorced
Clint	28	31-12-2000	Software Development	Undergraduate	Male	Single

To view, select a column in the table.

view as

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



Search for add rows, drag and drop in the canvas

Microsoft Azure Machine Learning Studio

Experiment created on 9/17/2017

In draft Properties Project

add ↗

Saved Datasets

- Samples
- Restaurant ratings

Data Transformation

- Manipulation
- Add Rows

Employee Dataset - AR1.csv Employee Dataset - AR2.csv

Add Rows

Prior Run

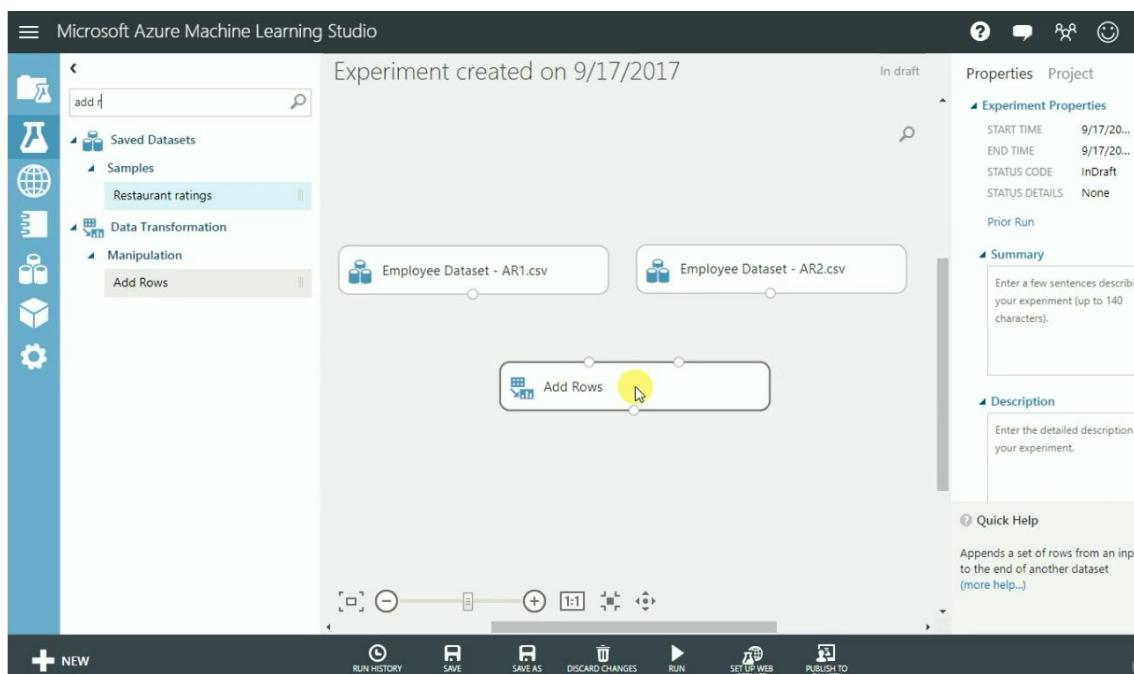
Summary

Description

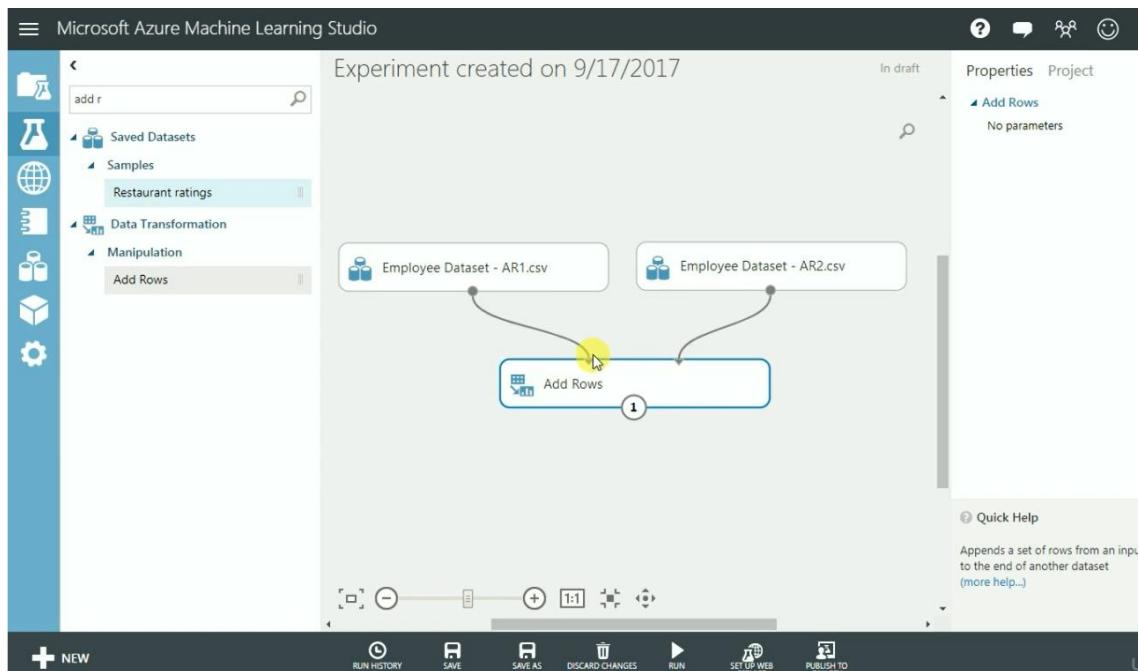
Quick Help

Appends a set of rows from an input to the end of another dataset (more help...)

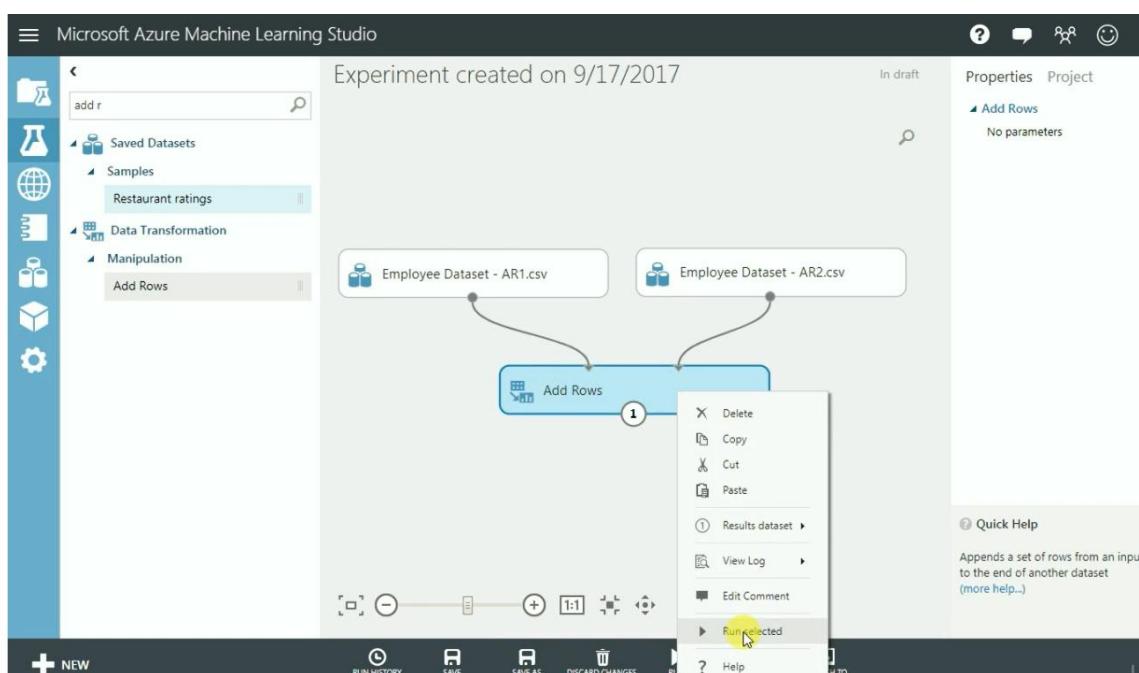
RUN HISTORY SAVE SAVE AS DISCARD CHANGES RUN SET UP WEB SERVICE PUBLISH TO GALLERY



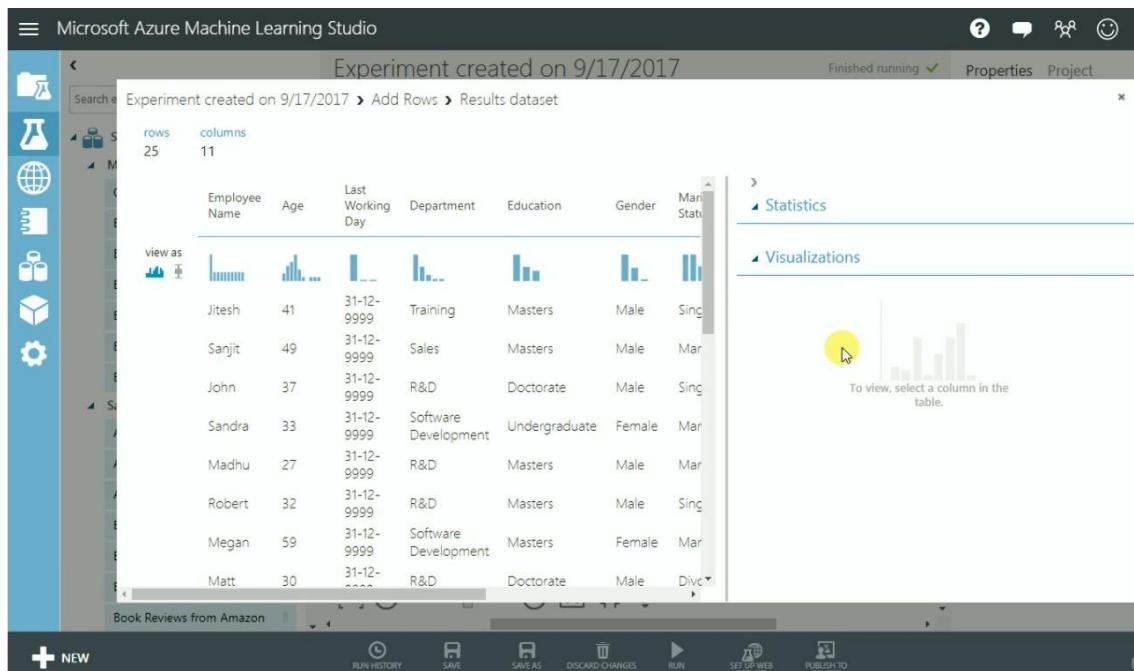
Connect both the output nodes from csv to Add row



## Run and execute



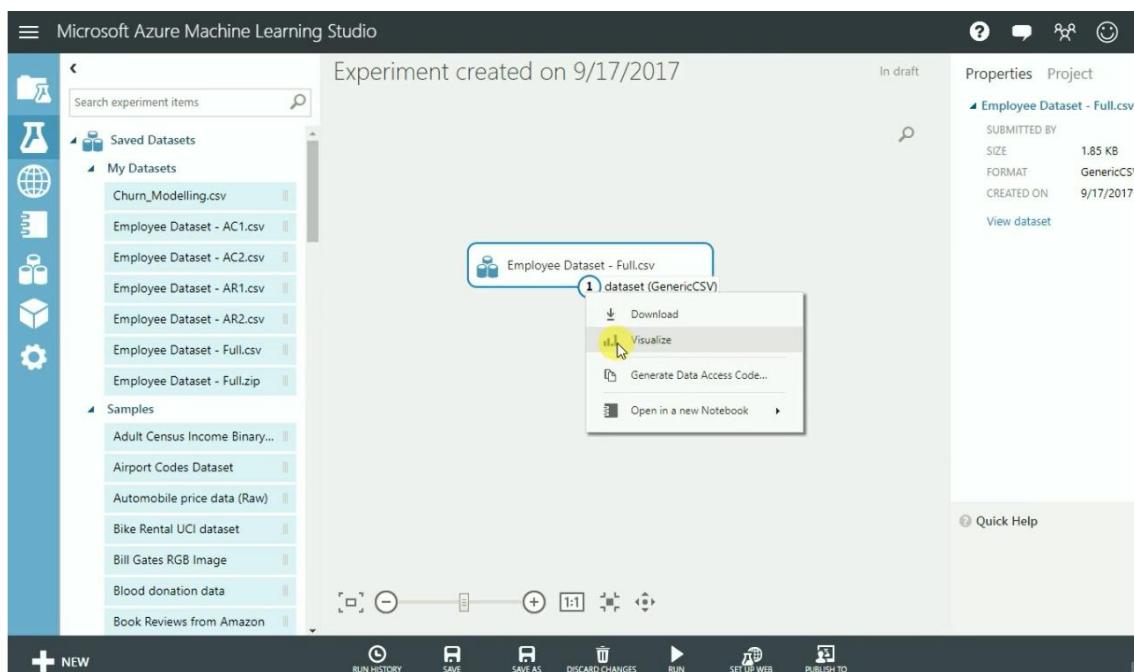
## Run and visualize the output



## Data Manipulation Using Remove Duplicate Components

### Remove Duplicate

Visualize a data set with duplication



Can see three duplications in the visualization

Experiment created on 9/17/2017 > Employee Dataset - Full.csv > dataset

rows 25 columns 11

	Will	George	Adam	Adam	Emma	Vin	Richard	Clint	Kate	Morgan
	38	36	35	35	29	31	34	28	29	32
	2012	31-12-9999	31-12-9999	31-12-9999	31-12-9999	31-12-9999	31-12-9999	31-12-9999	31-12-9999	31-12-9999
	R&D	Software Development								
	Doctorate	Masters	Masters	Masters	Undergraduate	Undergraduate	Undergraduate	Undergraduate	Undergraduate	Undergraduate
	Male	Mar	Male	Male	Female	Male	Male	Male	Female	Male
	Sing	Mar	Mar	Mar	Sing	Male	Divc	Sing	Divc	Divc

Search for remove duplicate rows, drag and drop the same in canvas

Experiment created on 9/17/2017

In draft

Properties Project

Experiment Properties

- START TIME 9/17/20...
- END TIME 9/17/20...
- STATUS CODE InDraft
- STATUS DETAILS None

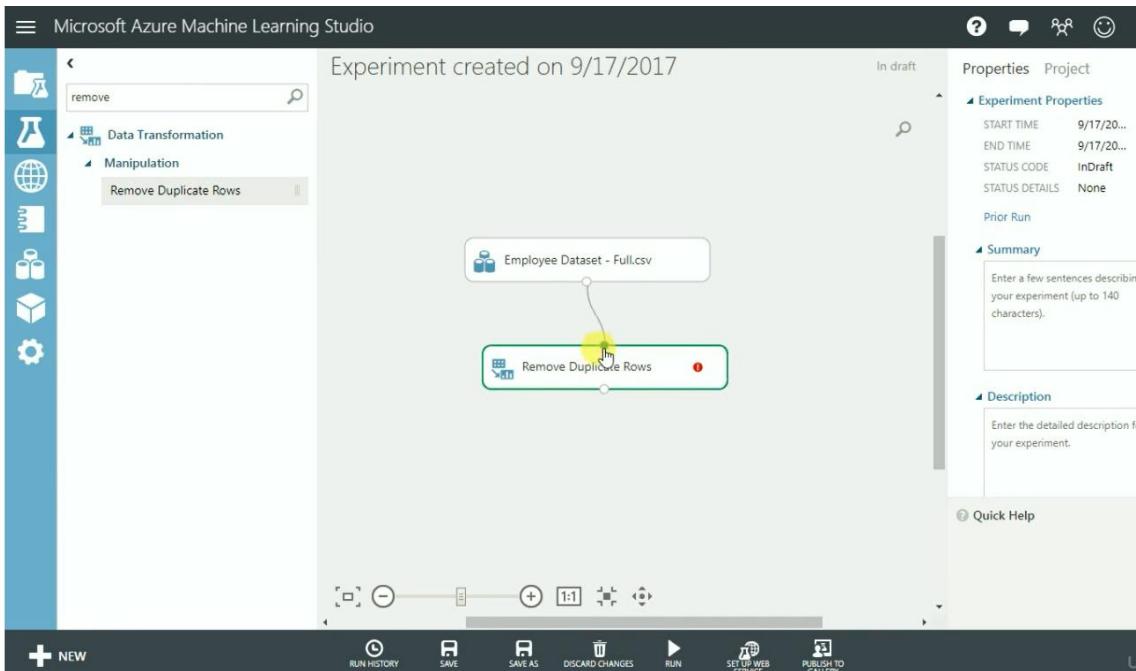
Prior Run

Summary

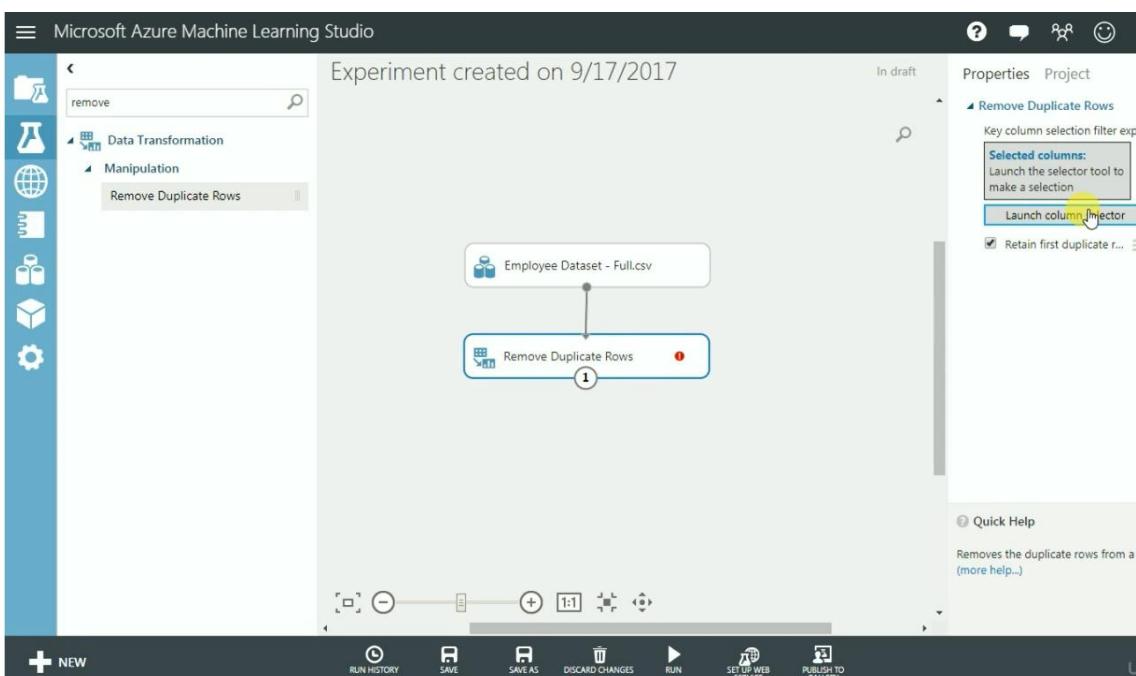
Description

Quick Help

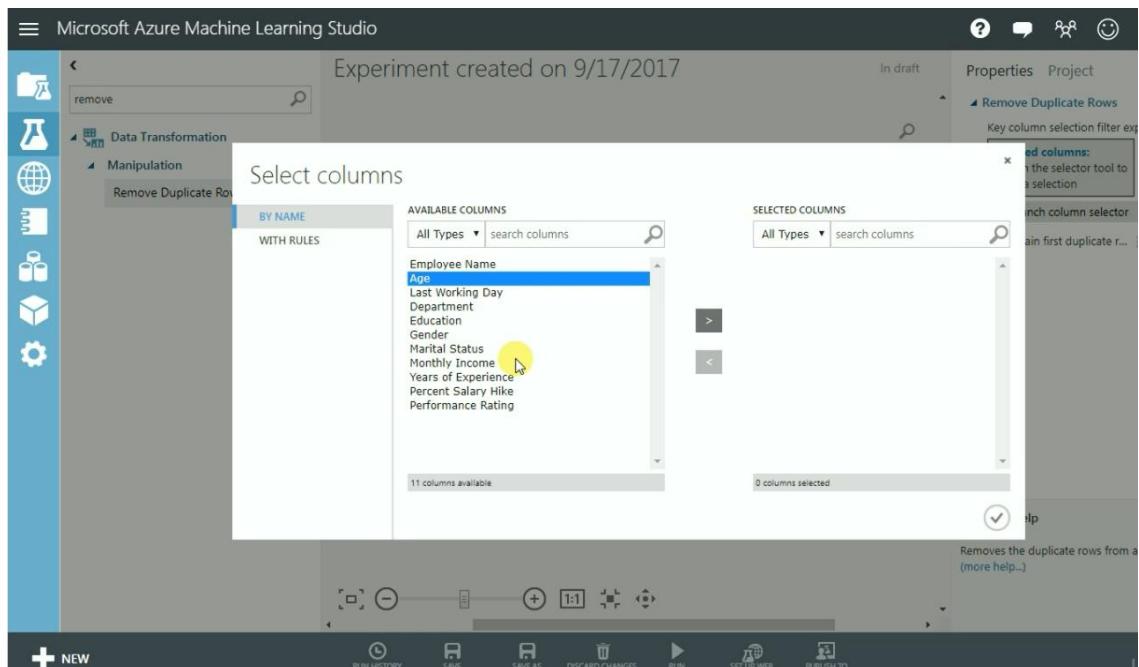
Connect both input and output nodes



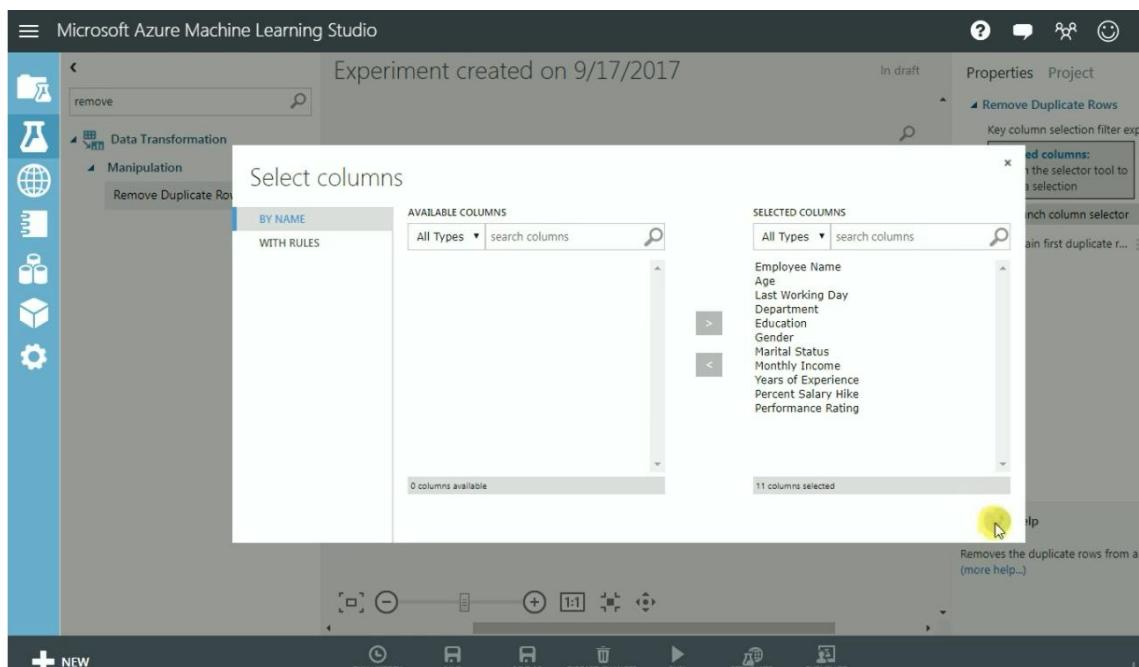
Click on Launch column selector



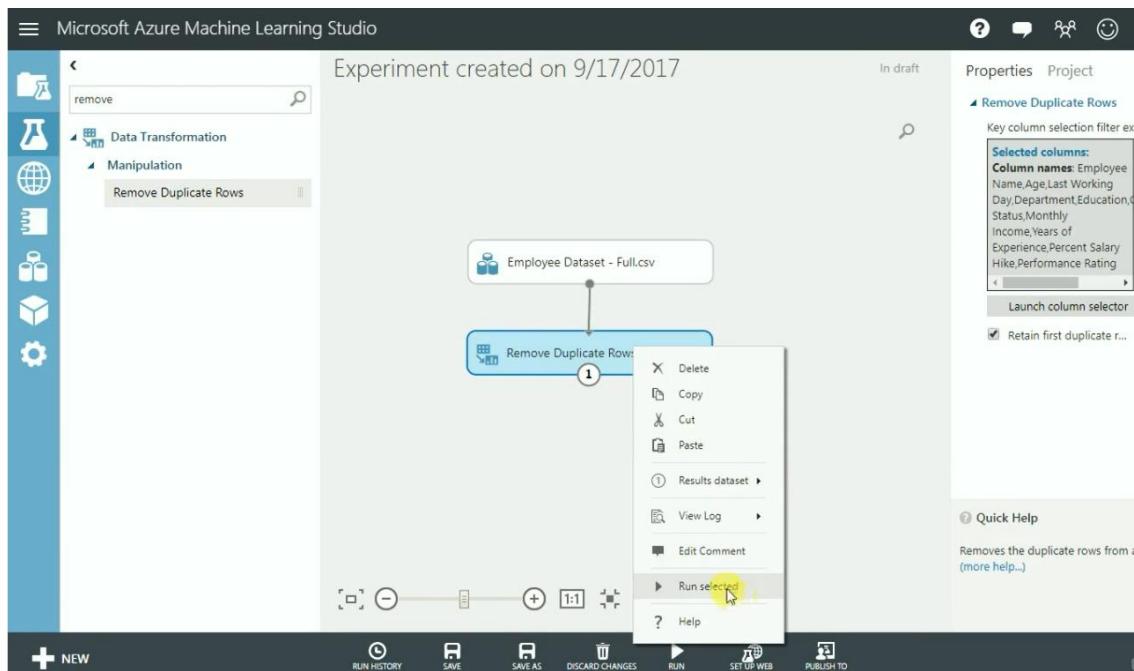
From the available columns select all the types and move to selected columns



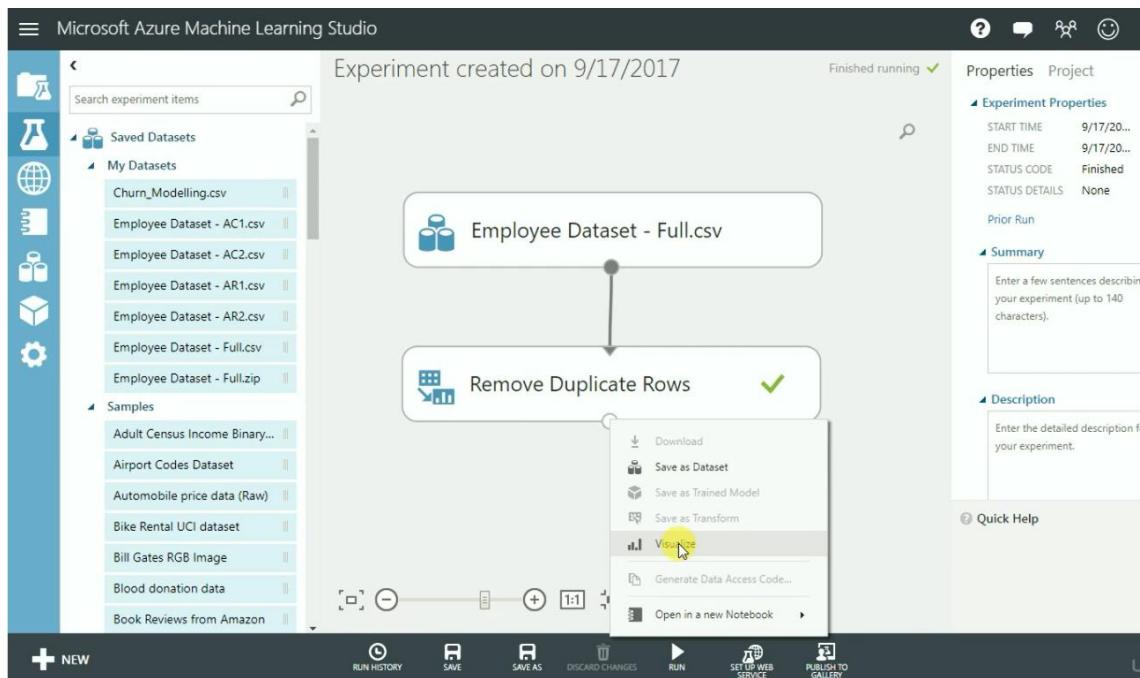
Select all columns and click ok



Run the module



## Visualize after successful execution



Can view that duplication removed successfully

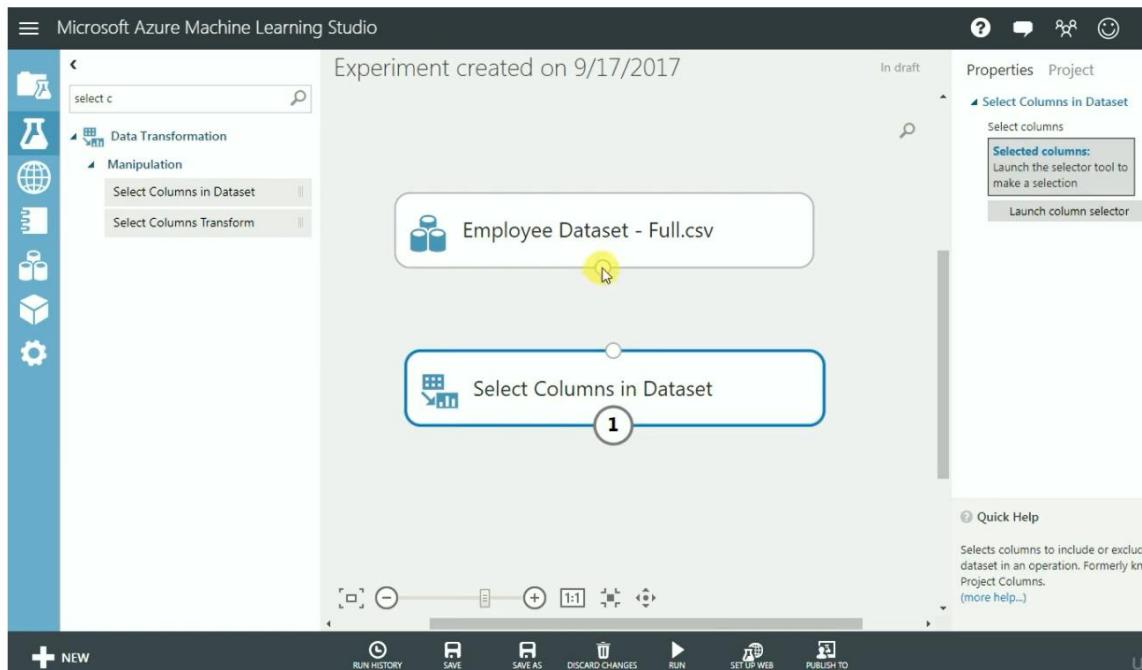
The screenshot shows the Microsoft Azure Machine Learning Studio interface. In the center, there is a table titled "Experiment created on 9/17/2017" with the path "Remove Duplicate Rows > Results dataset". The table has 23 rows and 11 columns. A yellow circle highlights the row for "George". The columns include Name, Age, Date, Department, Education, Gender, and Marital Status. The "Statistics" and "Visualizations" sections are visible on the right side.

## Data Manipulation Using Select Column in a Dataset Component

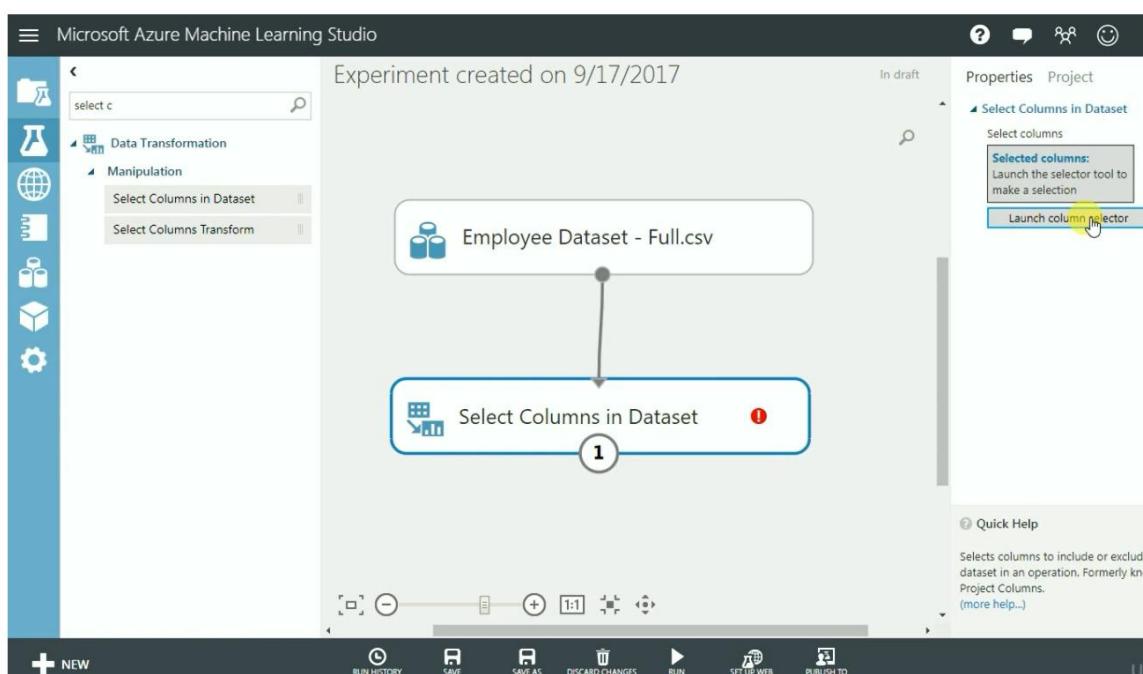
Take a dataset which is uploaded already in canvas

The screenshot shows the Microsoft Azure Machine Learning Studio interface with the experiment canvas. A dataset component, labeled "Employee Dataset - Full.csv", is highlighted with a yellow circle. The canvas includes various components like "Run History", "Save", "Discard Changes", "Run", "Set Up Web", and "Publish To". On the left, there is a sidebar with "Saved Datasets" and "Samples" sections. On the right, there are sections for "Properties", "Experiment Properties", "Summary", and "Description".

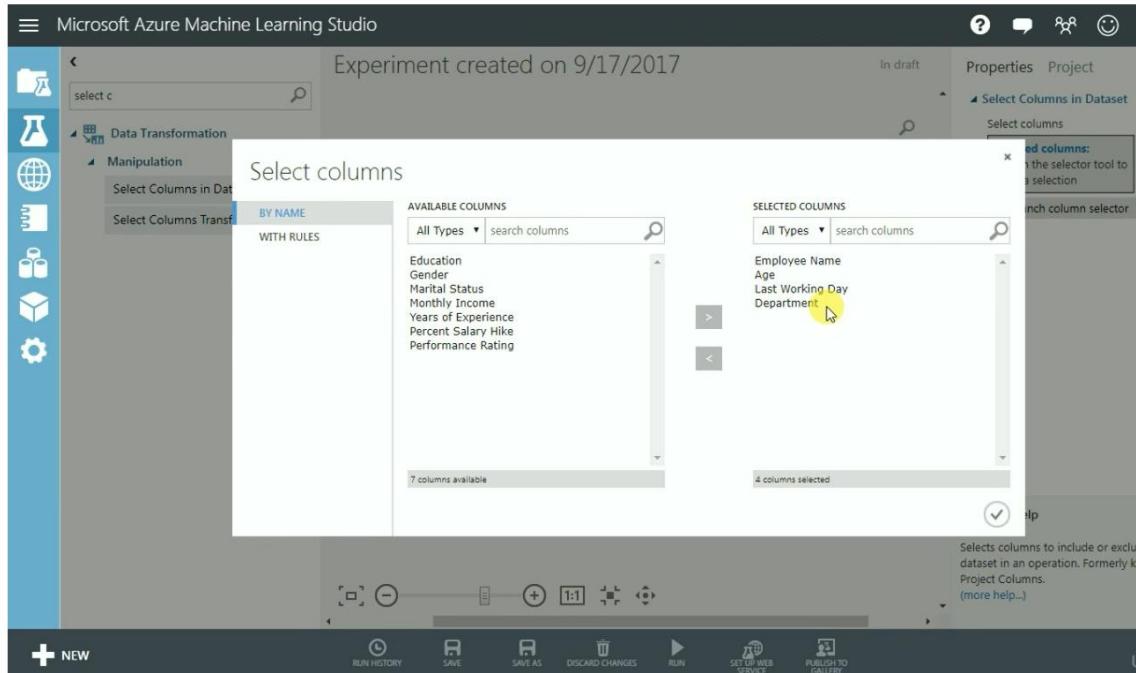
Search for select columns for datasets, drag and drop in canvas



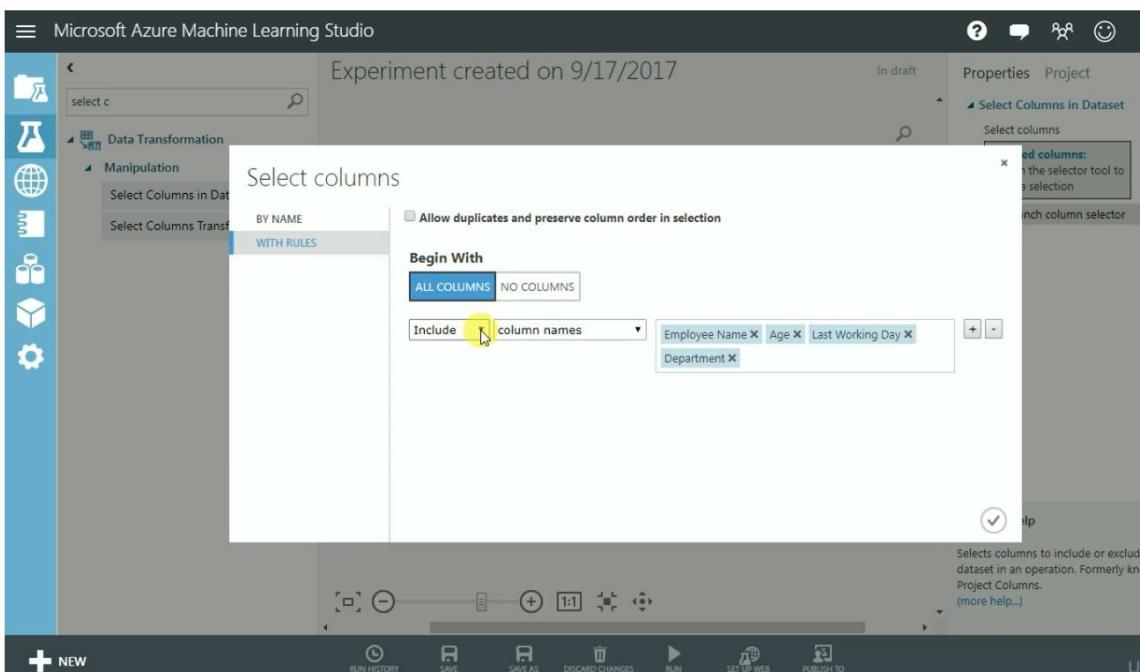
Connect both output and input nodes and click on launch column sector



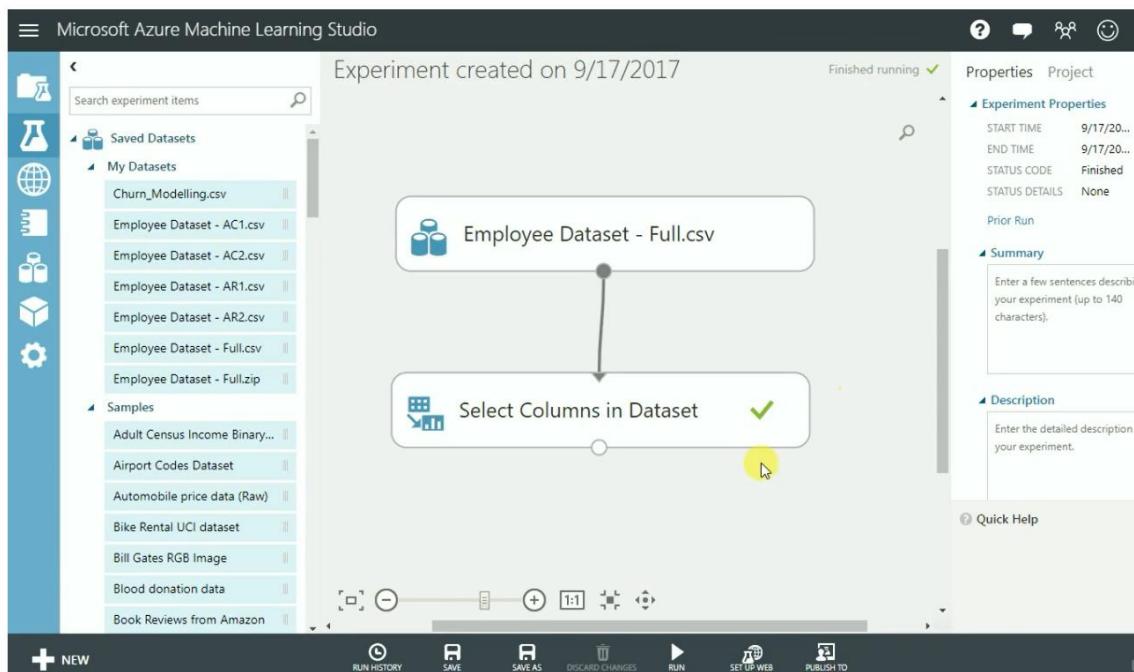
Select first four types and move to selected columns as below and press ok



By clicking (with rules) can select any types as per requirement



Run the module after selection and visualize the output

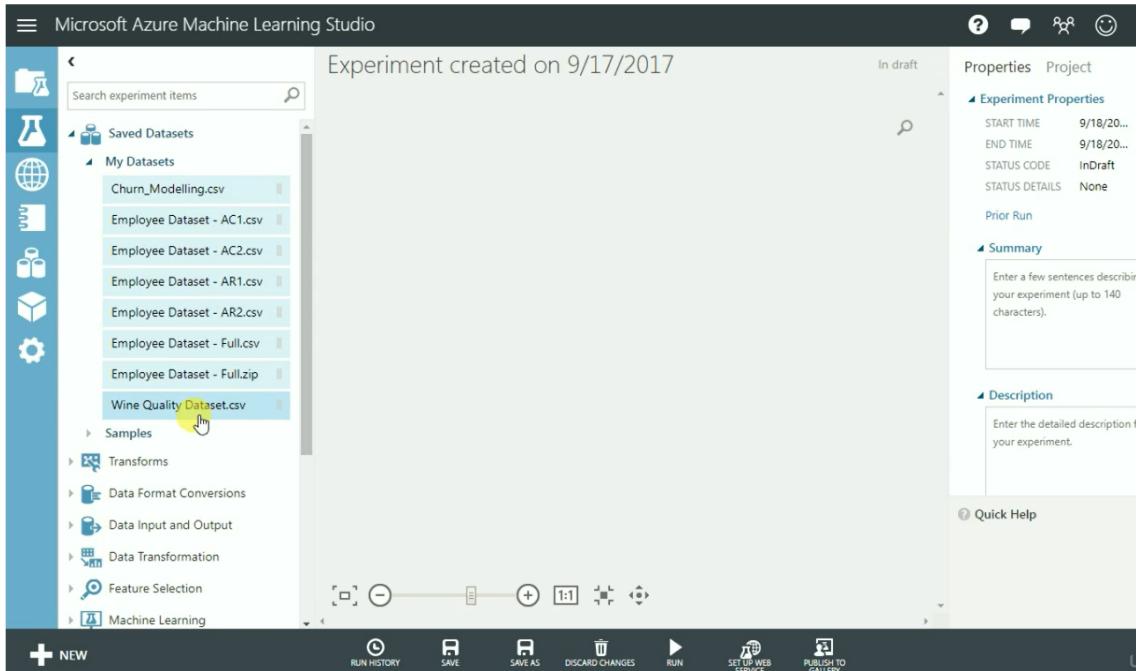


As per requirement output obtained for selected four columns successfully

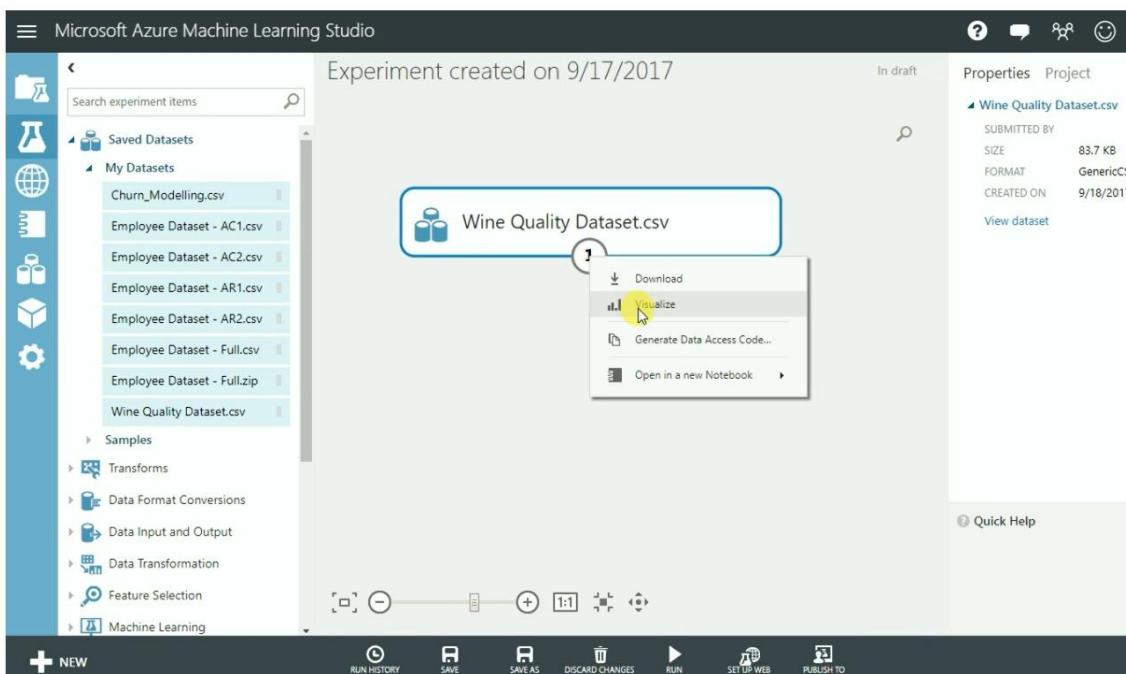
	Employee Name	Age	Last Working Day	Department
1	Jitesh	41	31-12-9999	Training
2	Sanjit	49	31-12-9999	Sales
3	John	37	31-12-9999	R&D
4	Sandra	33	31-12-9999	Software Development
5	Madhu	27	31-12-9999	R&D
6	Robert	32	31-12-9999	R&D
7	Megan	59	31-12-9999	Software Development
8	Matt	30	31-12-9999	R&D
9	Will	38	01-03-2012	R&D
10	George	36	31-12-9999	Software Development
11	Adam	35	31-12-9999	Software Development
12	Adam	35	31-12-9999	Software Development
13	Adam	35	31-12-9999	Software Development

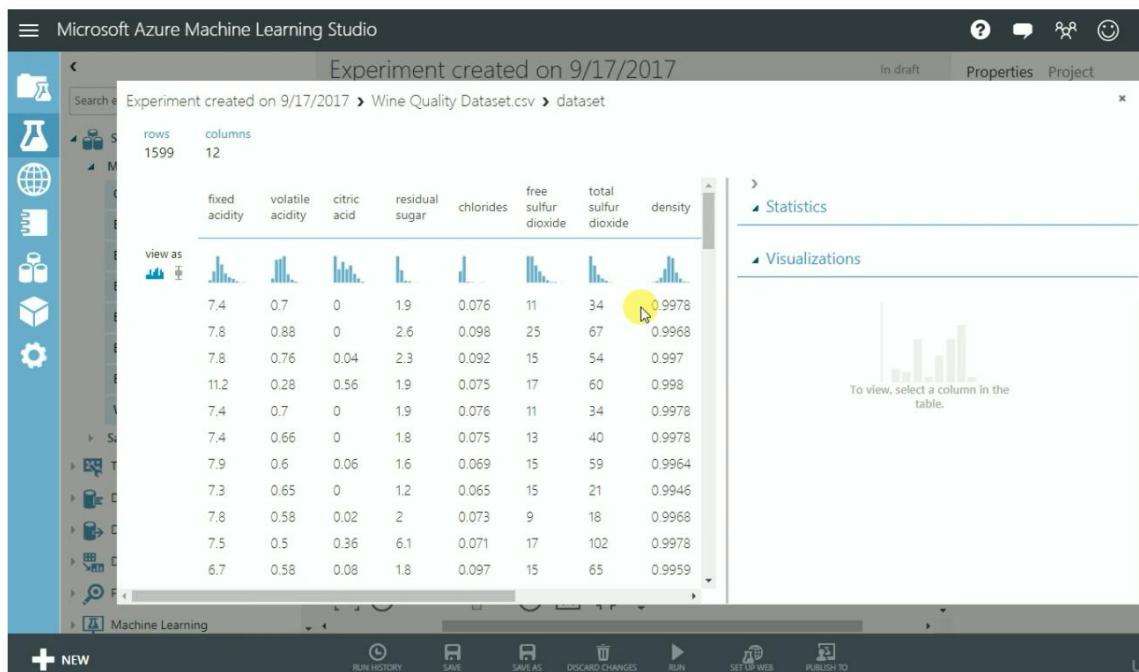
## Data Manipulation Using Apply SQL Transformation Component

Select the dataset uploaded

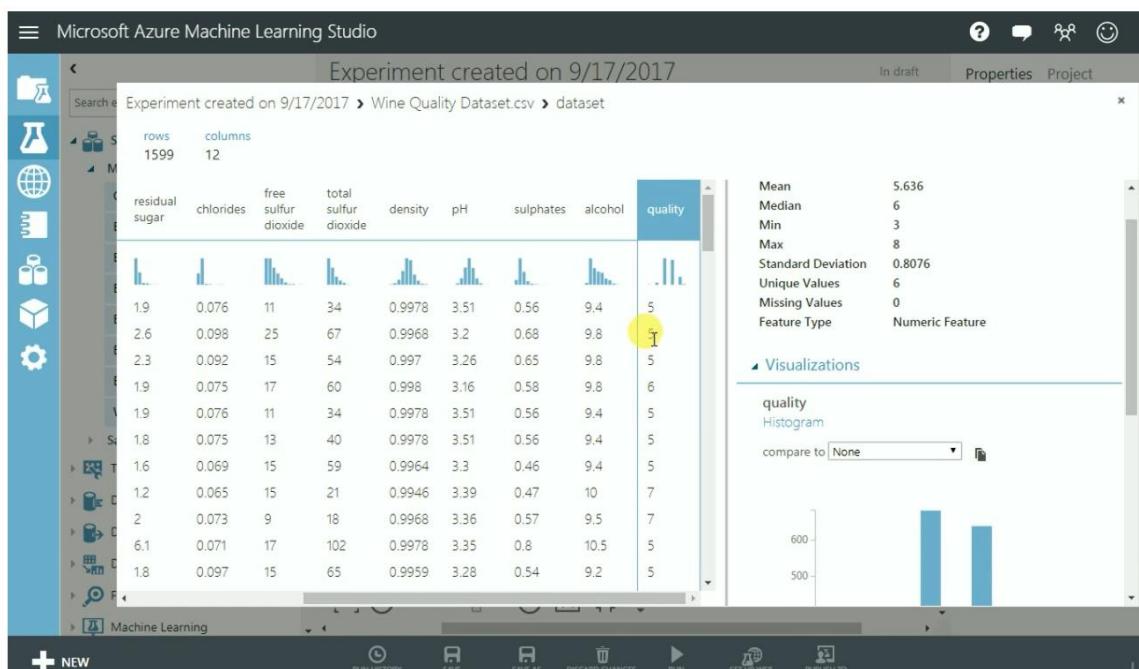


Drag and drop the selected data set and visualize the same

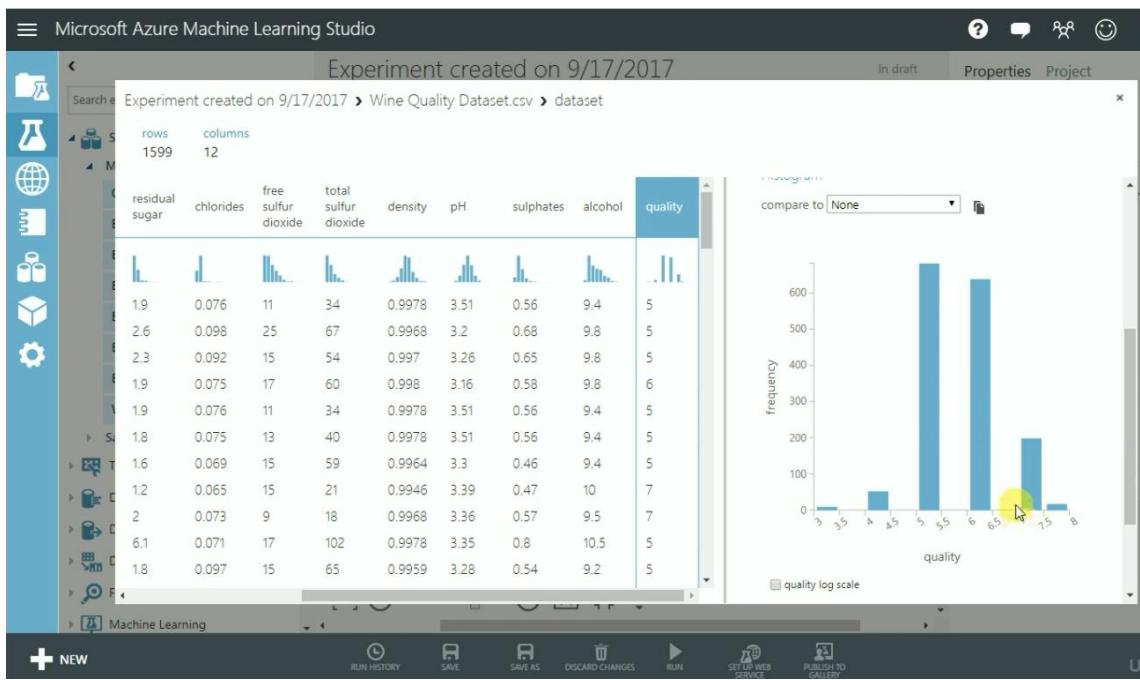




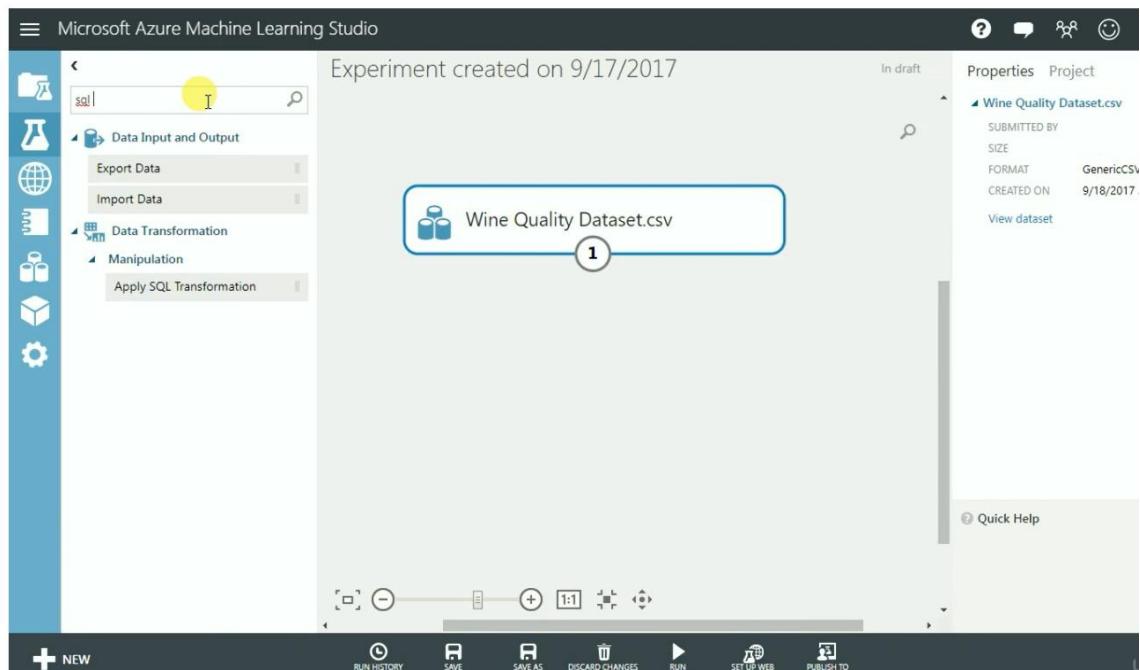
This dataset used for checking quality of wine, we can check the classification High, medium and low



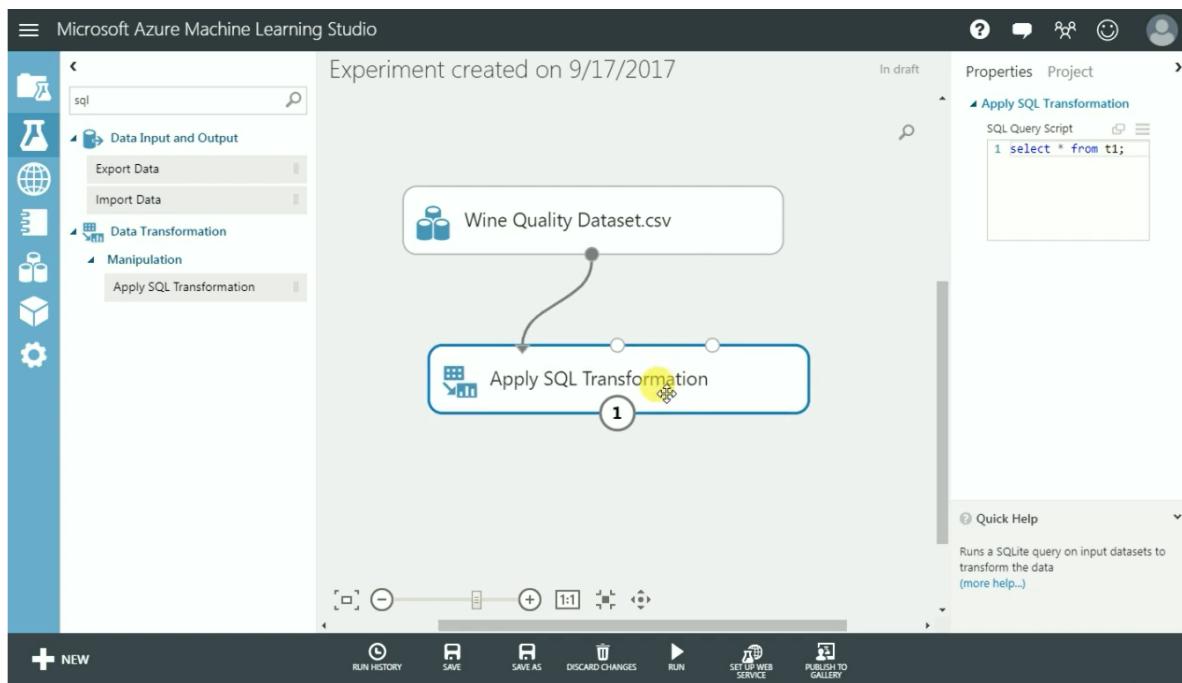
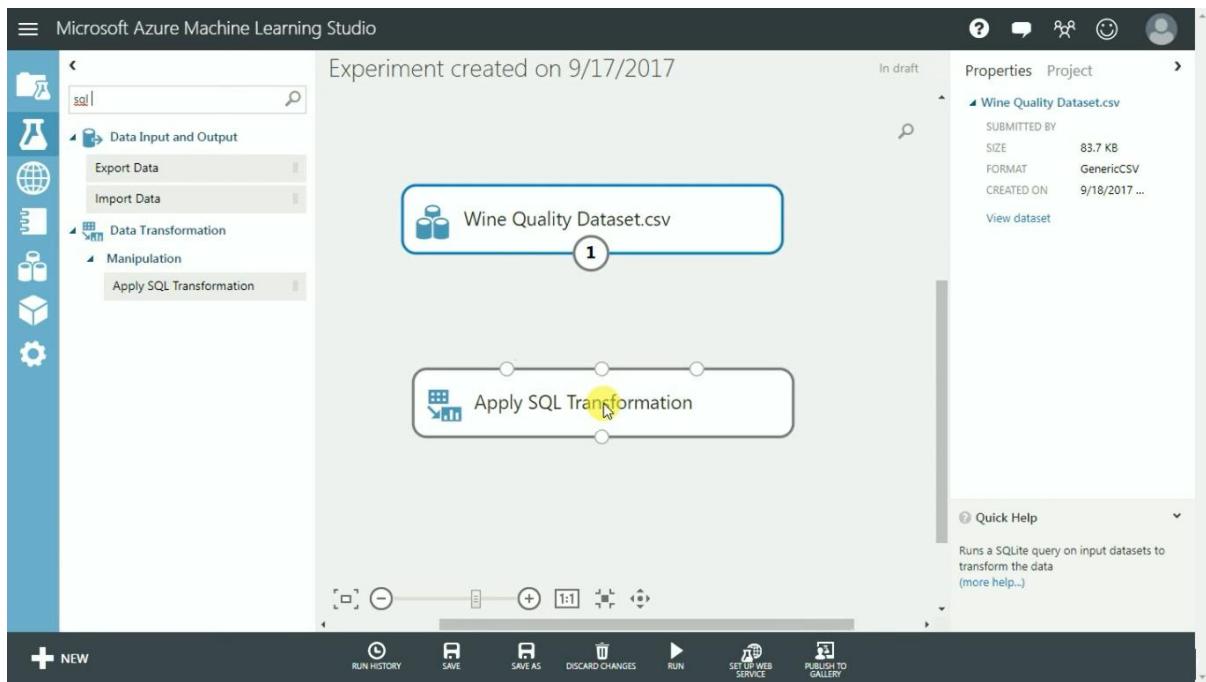
Considering 3 and 4 as low, 5 and 6 as medium, 7 and 8 as high from graph



## Search for Apply SQL transformation



## Connect the nodes



Enter the query script in the right to categorize

**SELECT \*,**

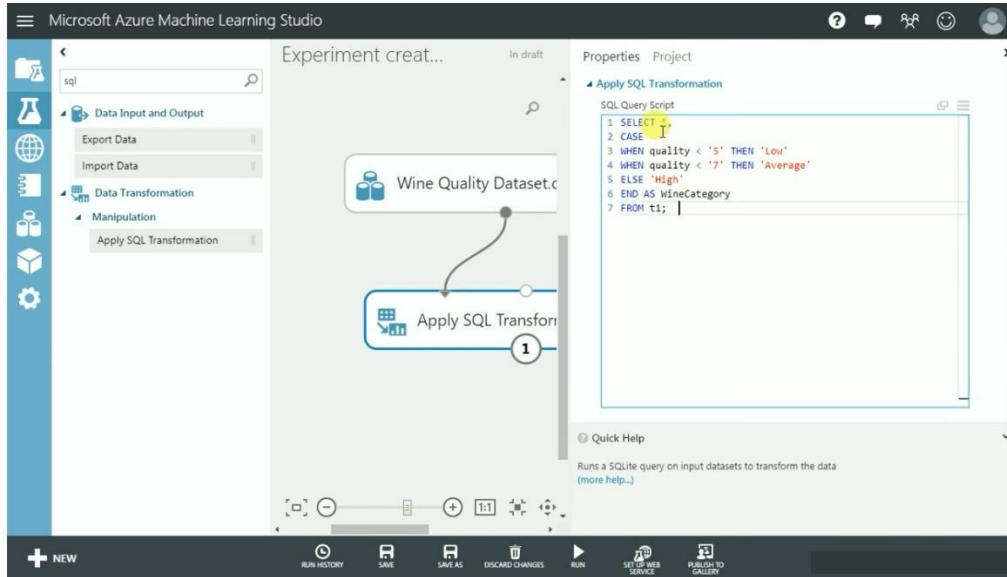
**CASE**

**WHEN quality < '5' THEN 'Low'**

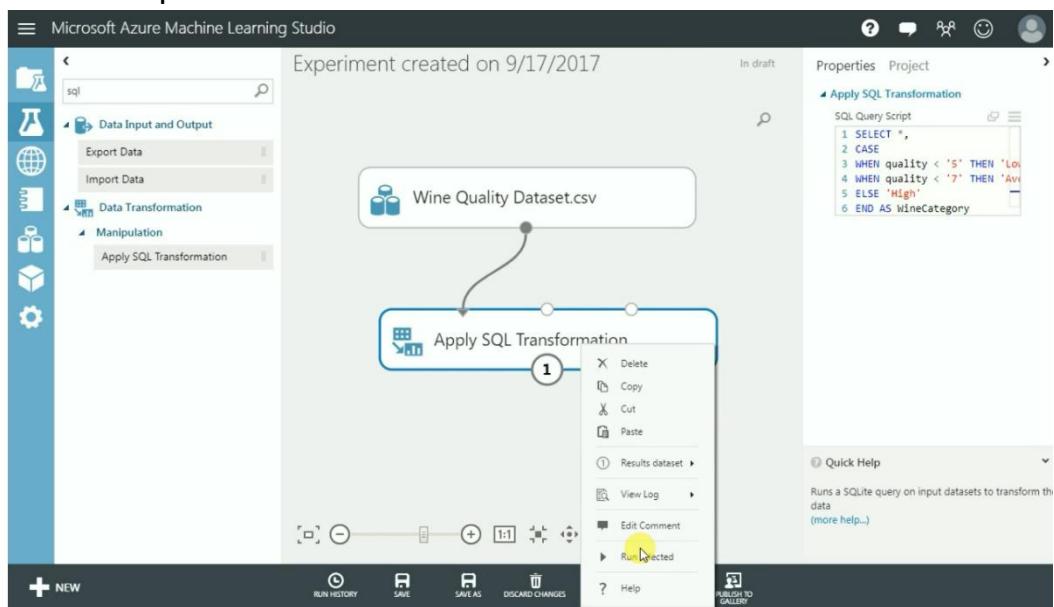
**WHEN quality < '7' THEN 'Average'**

ELSE 'High' END AS WineCategory

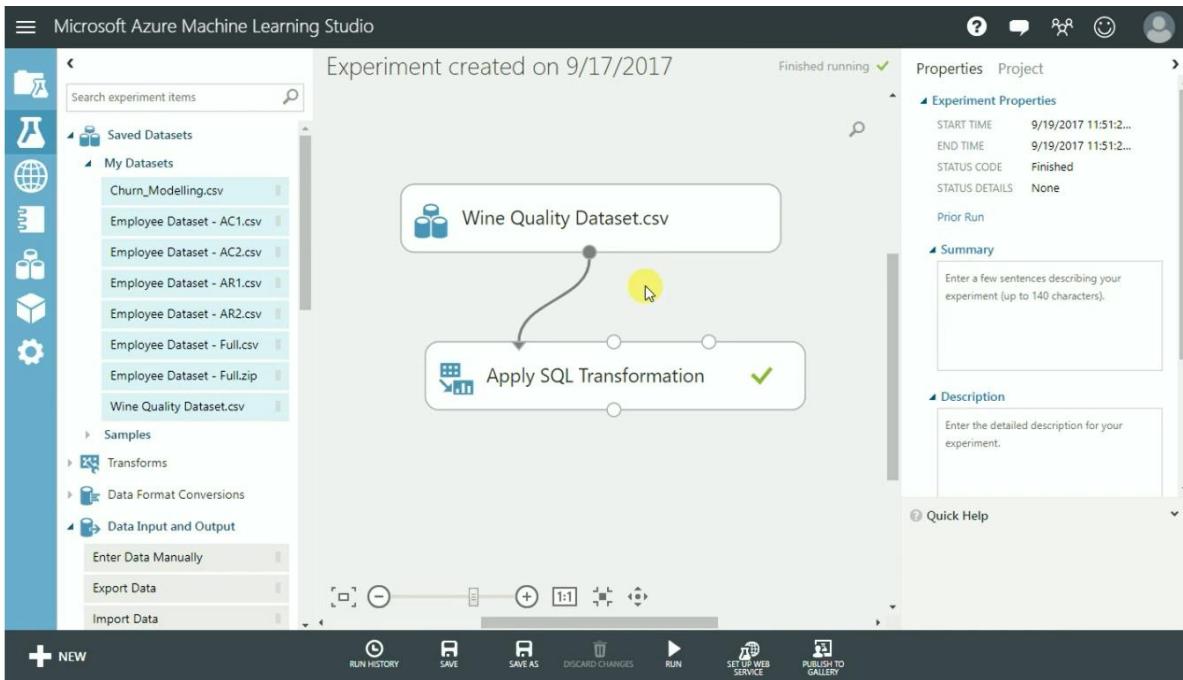
FROM t1;



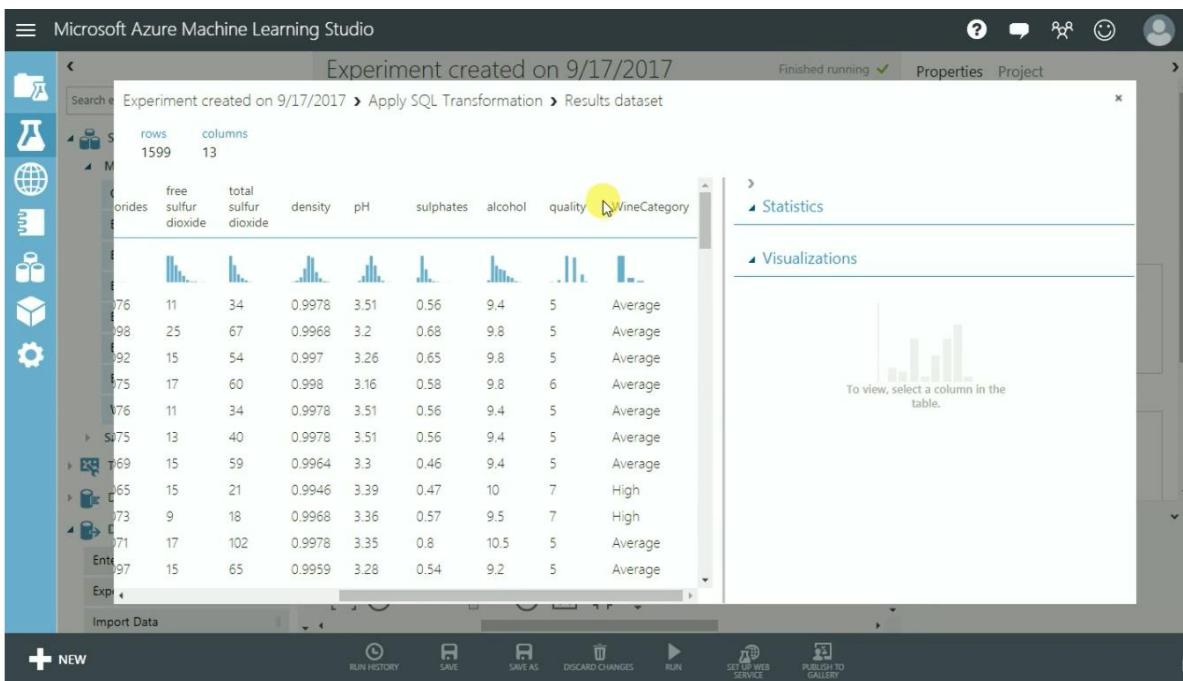
Run for output



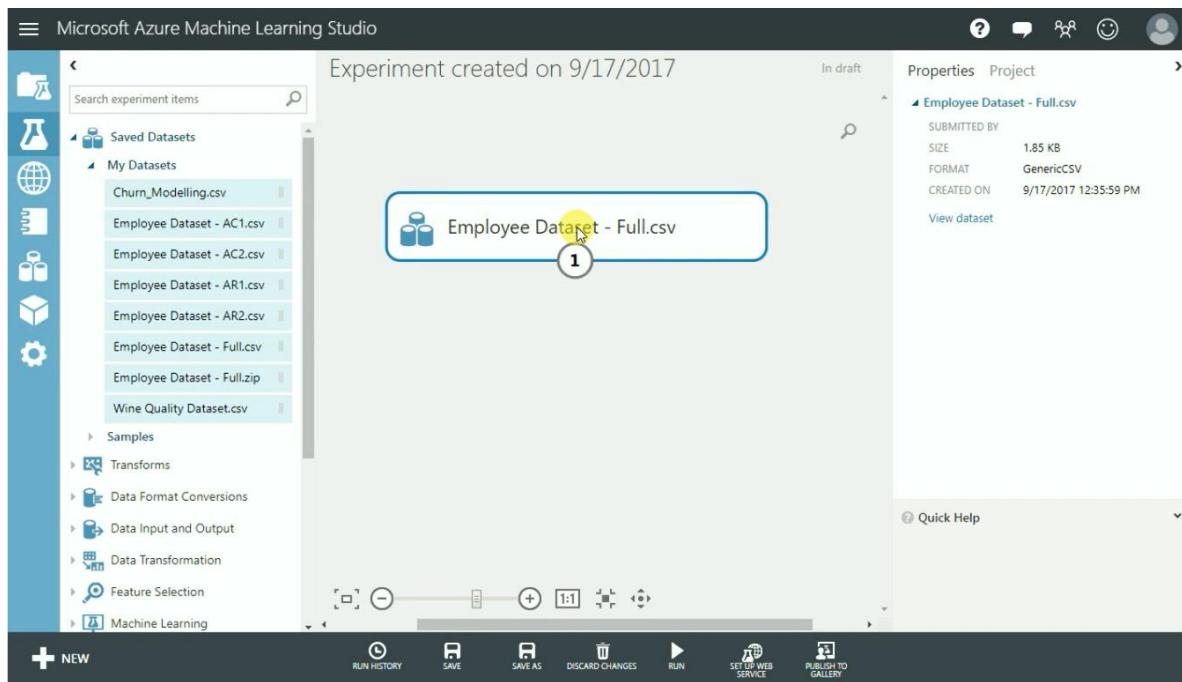
Right click and visualize for result



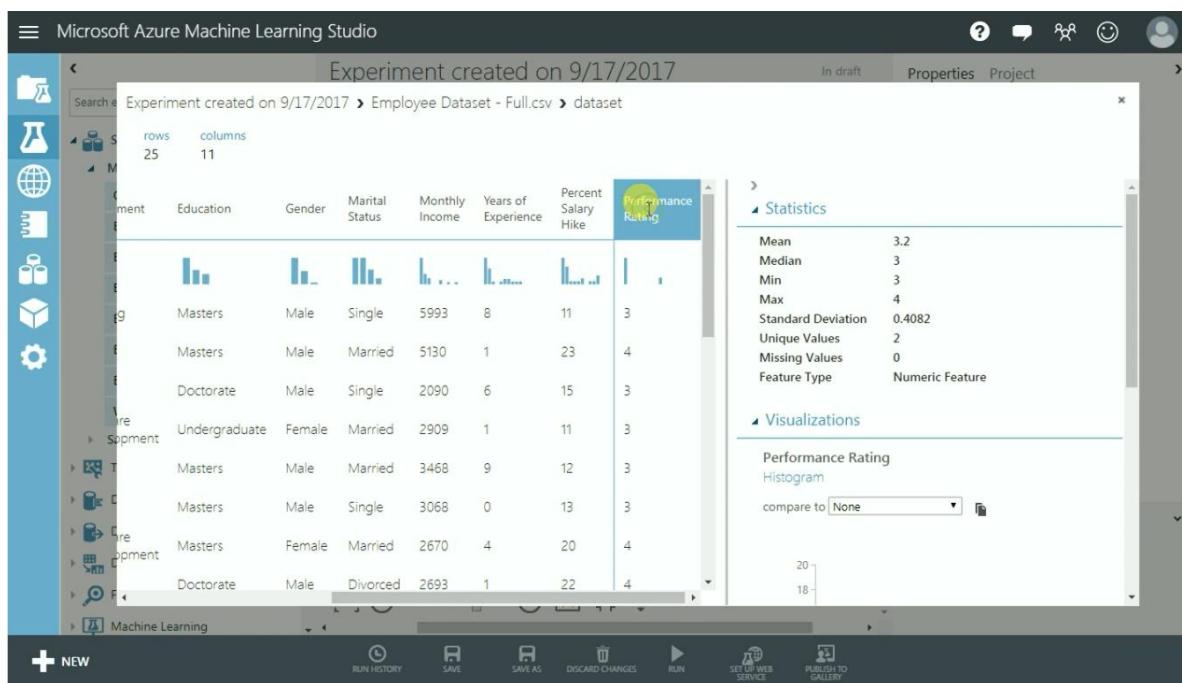
Hence obtained the result successfully using SQL transformation



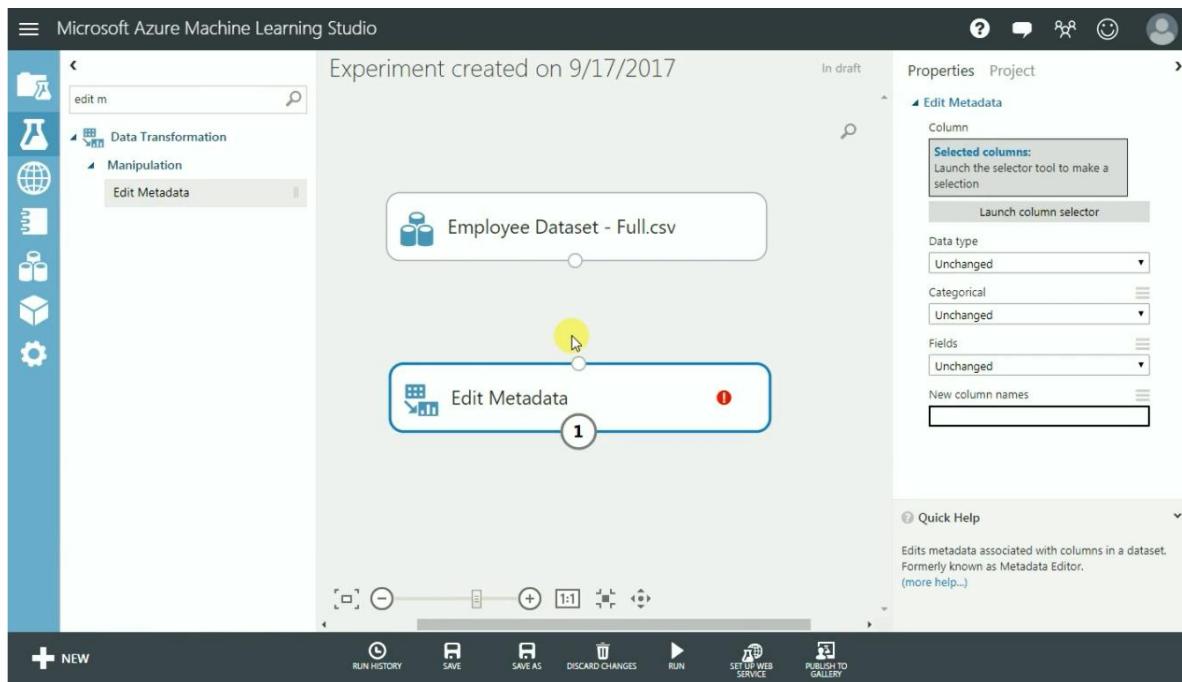
**Data Manipulation Using Edit Metadata Component**  
Select and drop the uploaded dataset and visualize



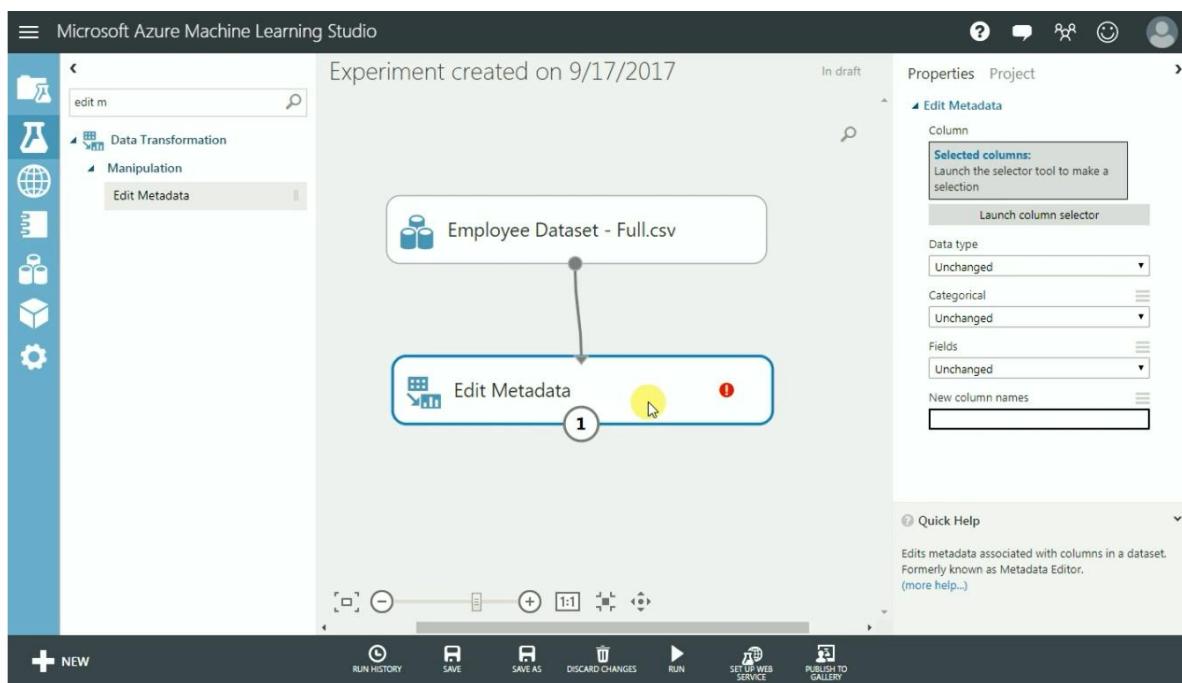
Using metadata can change numeric to category as per requirement



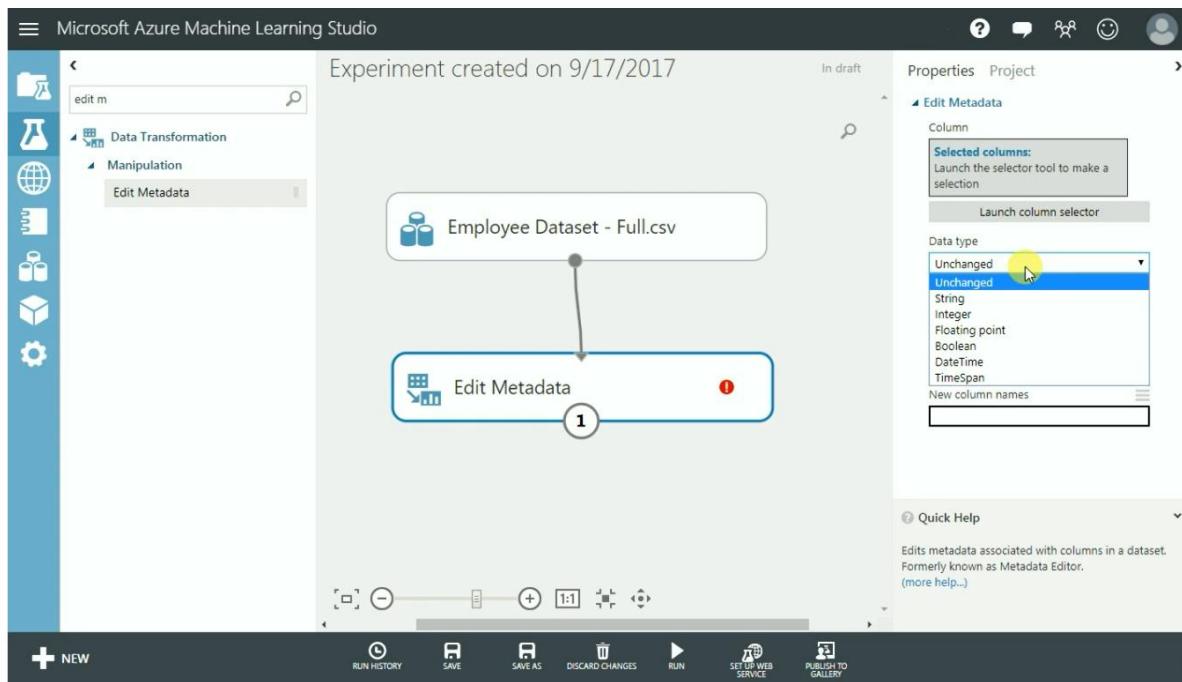
Search for Edit metadata and drop it in the canvas



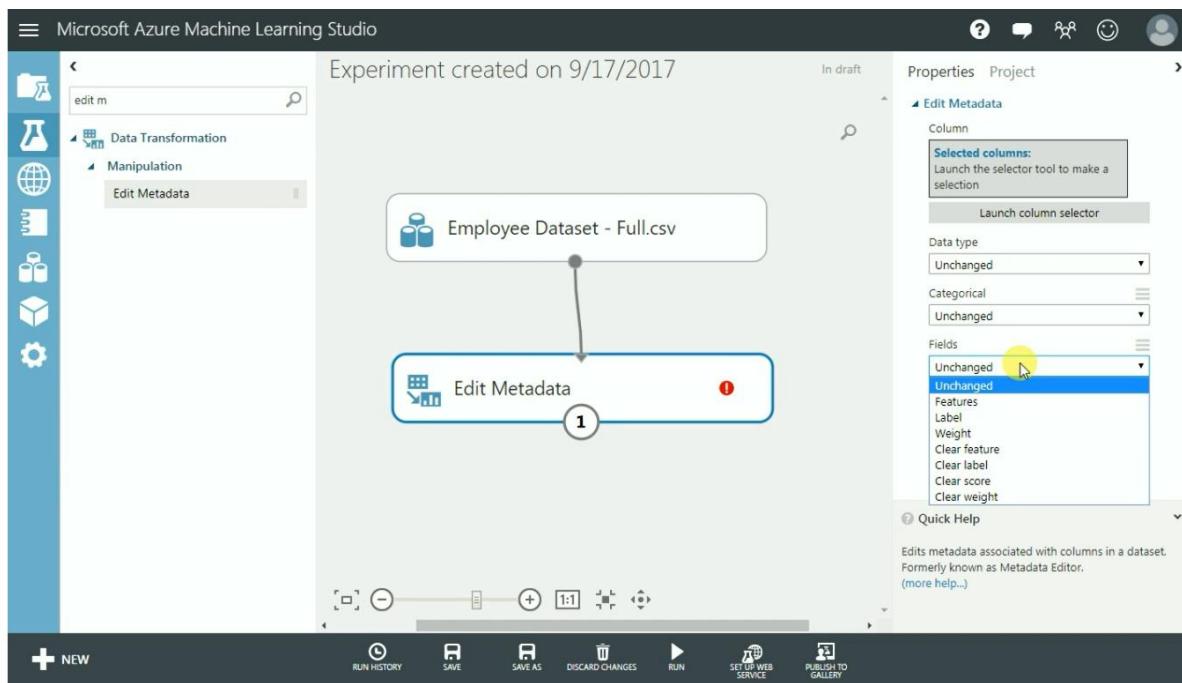
Attach both the data sets



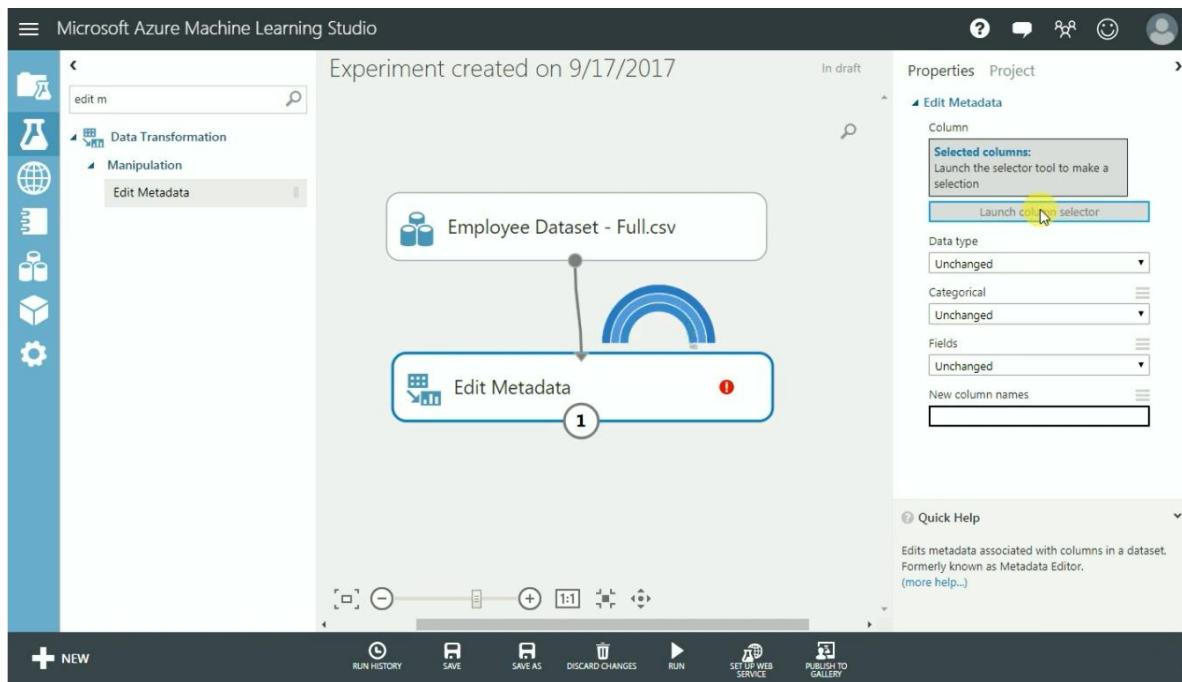
Can change supported data types as per requirement



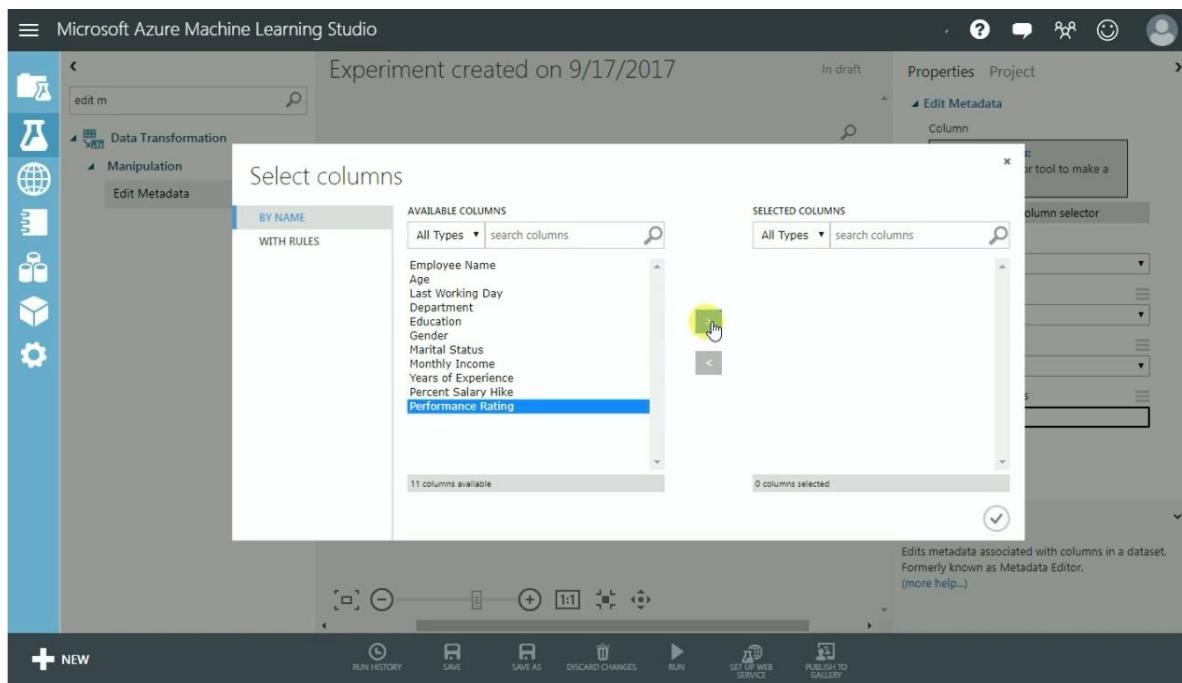
Also can change category and fields in the dropdown



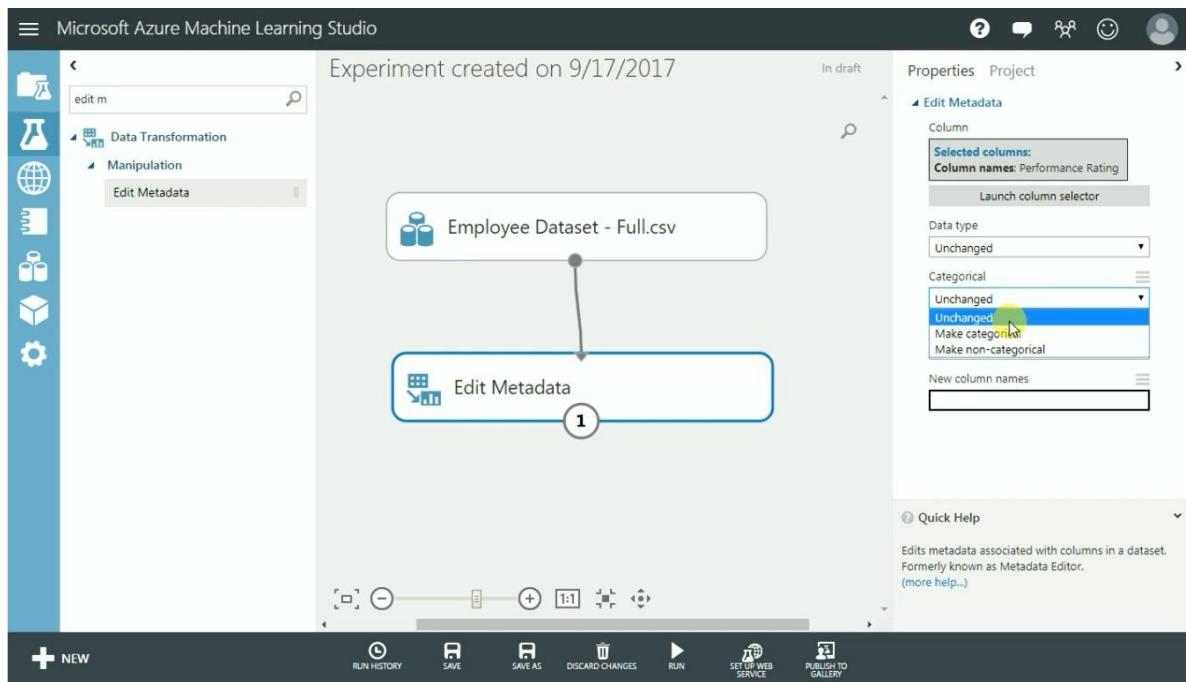
Click launch column sector to select the column in which metadata to be changed



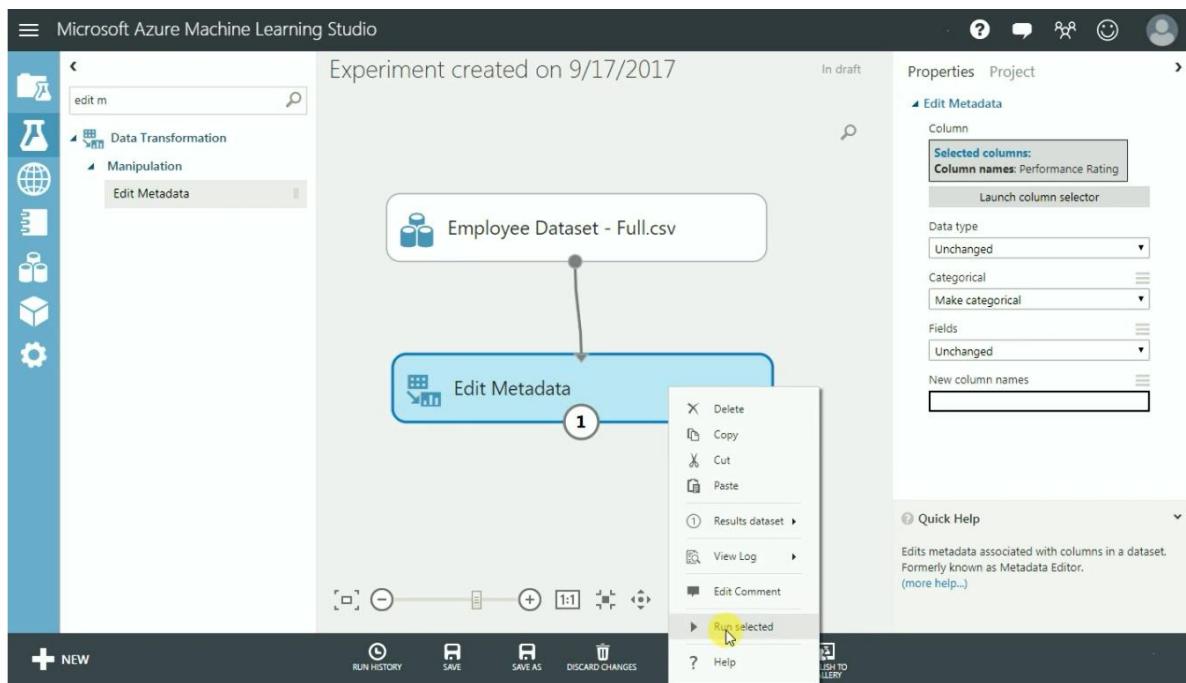
In this case select performance rating to change the metadata and click ok



Select make categorical from dropdown

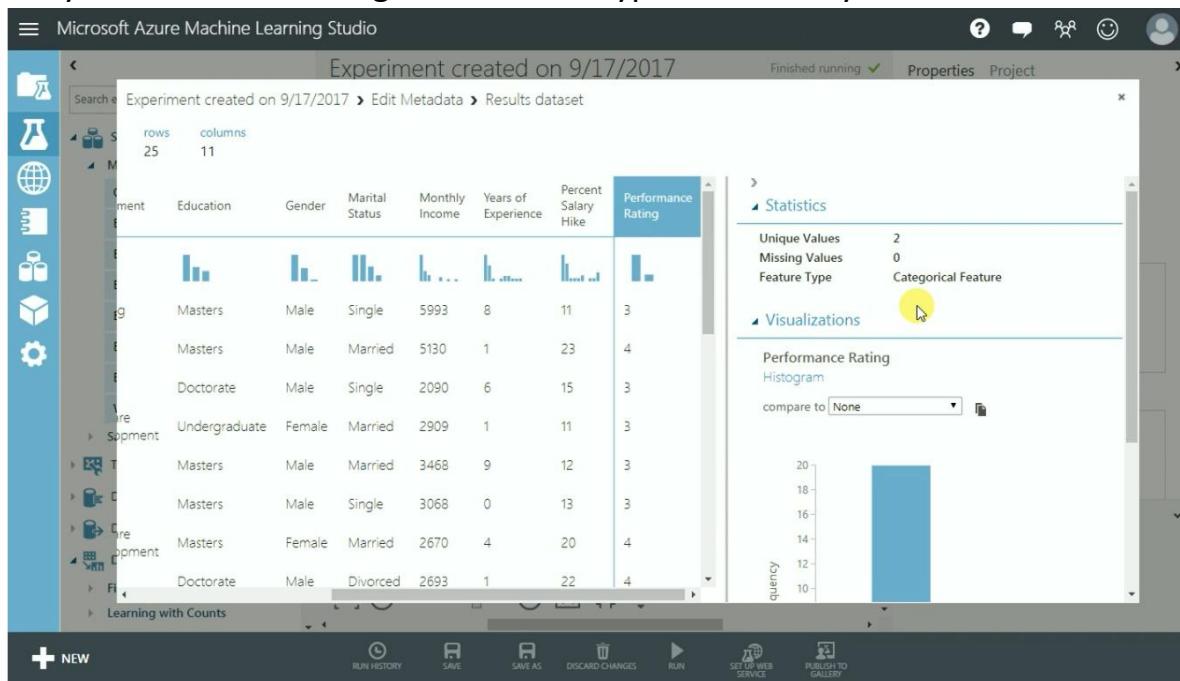


Run for obtaining result and click visualize after completion



Go to our column performance rating and select

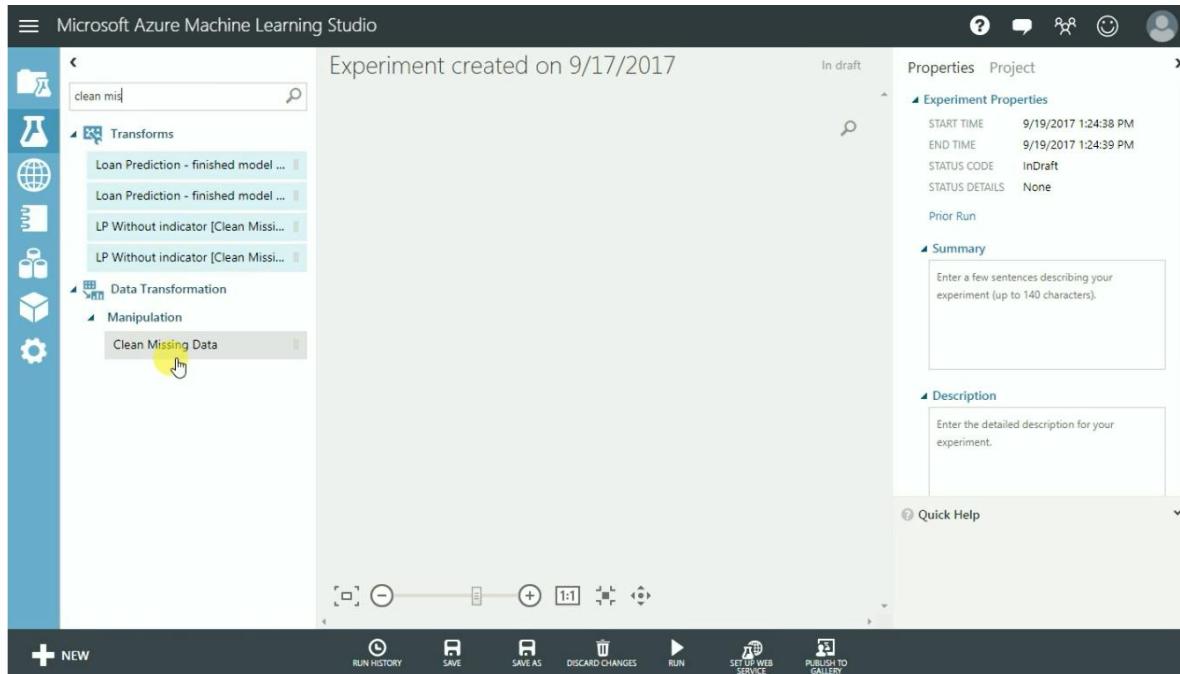
Now you can see the categorical feature type successfully



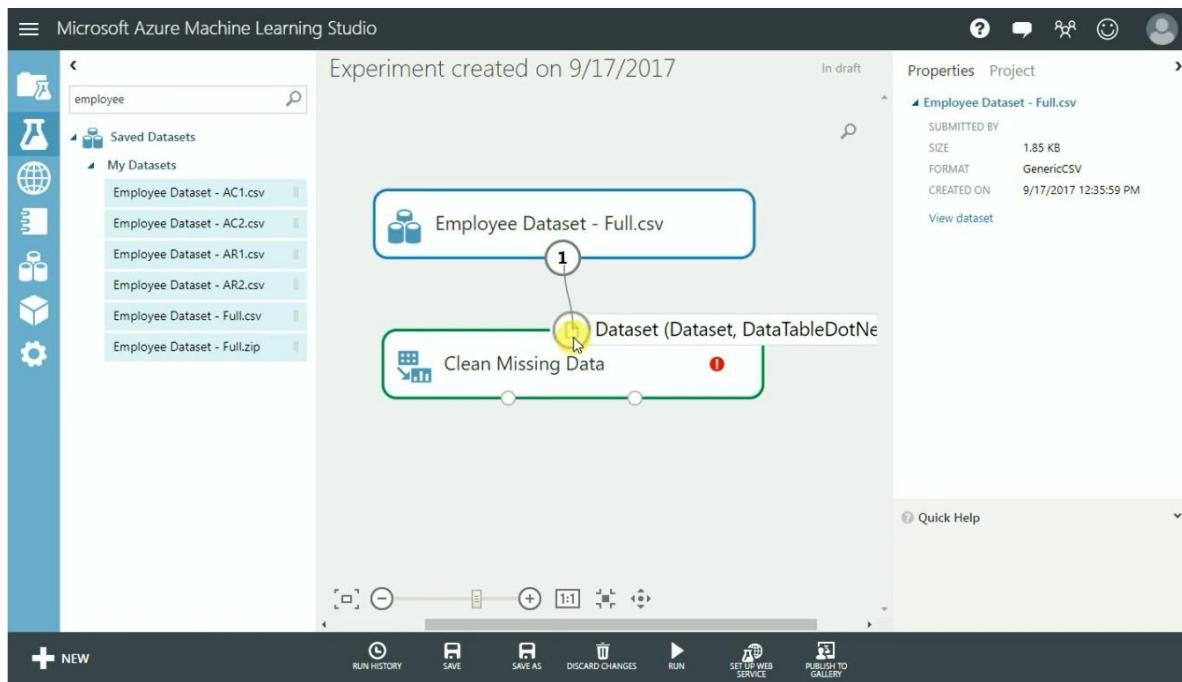
## Data Manipulation Using Clean Missing Data Component

### CLEAN MISSING DATA

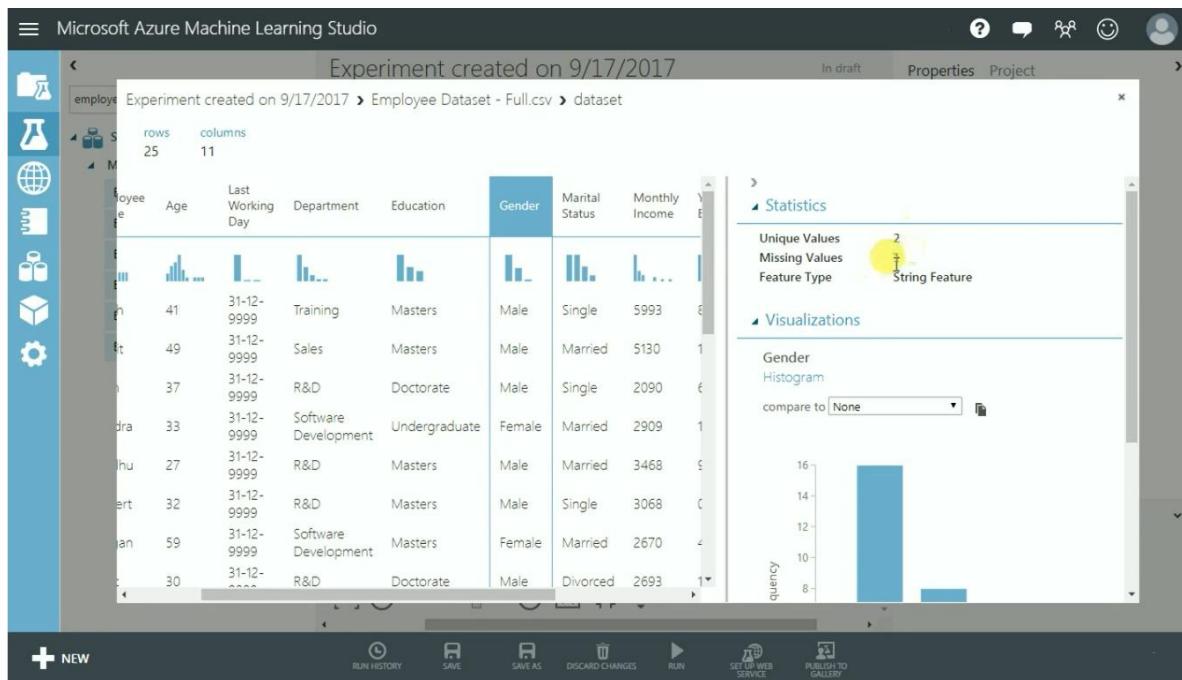
Search for clean missing data and drop in the canvas



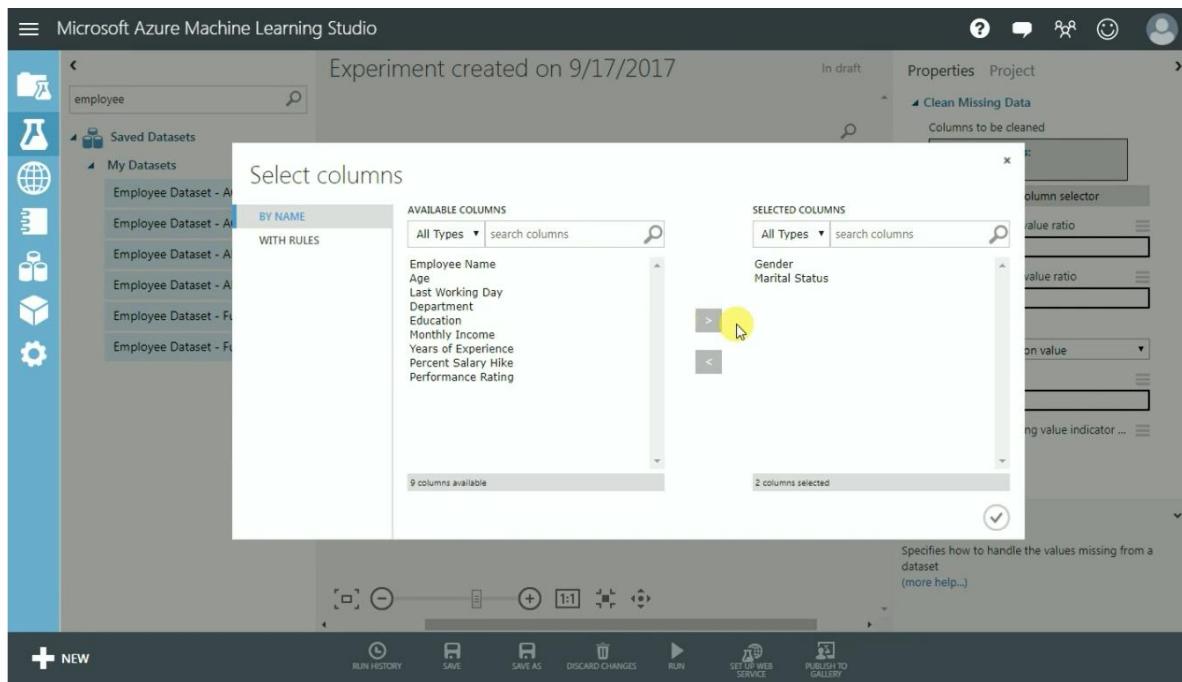
Also take the inputted dataset into canvas and attach with clean missing data



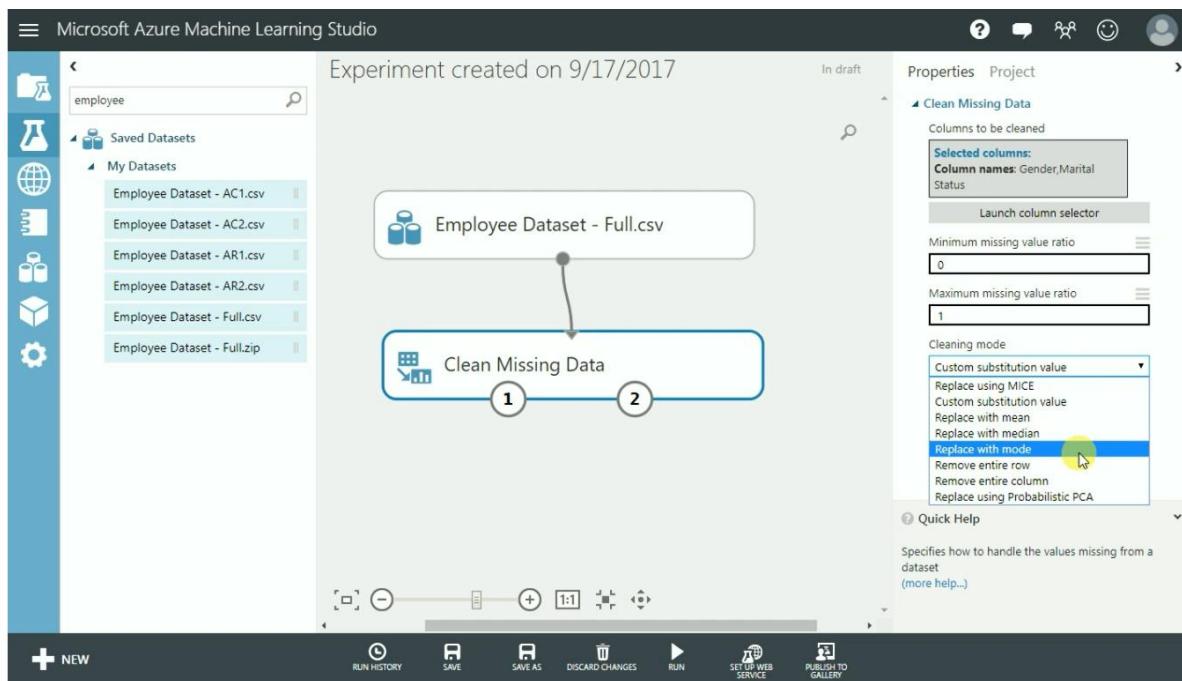
Visualize the dataset and check for any missing values in each columns



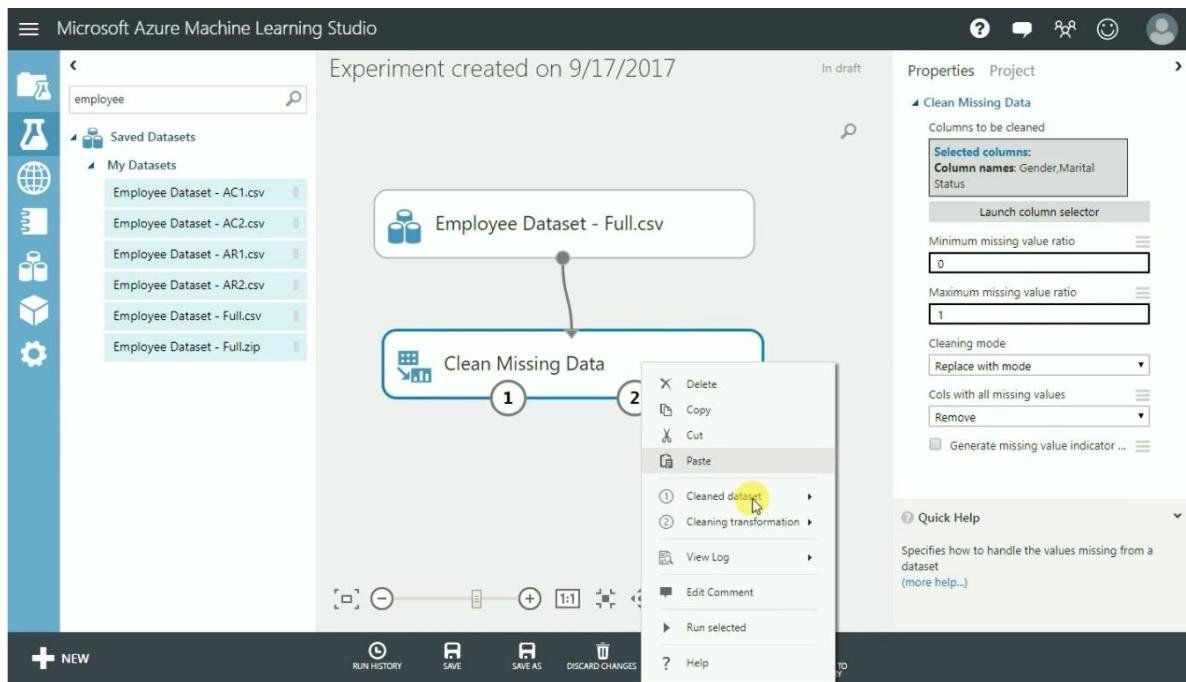
Launch column selector and select the two columns which has missing values



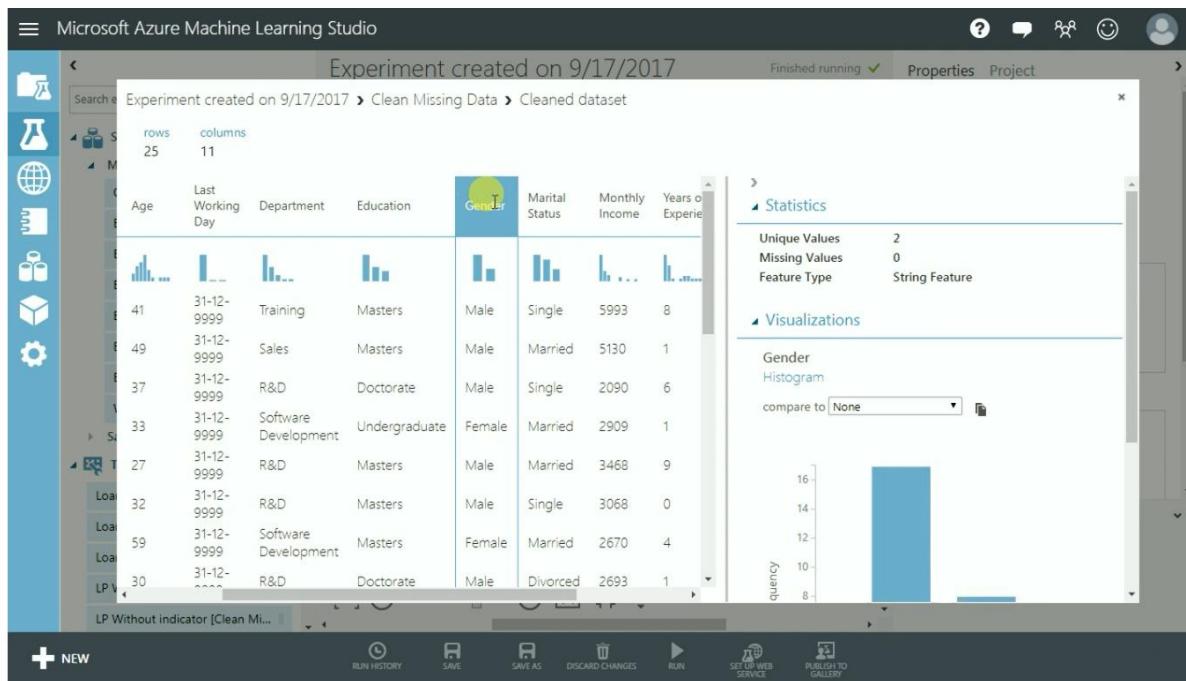
Select Replace with mode from dropdown



Run the module

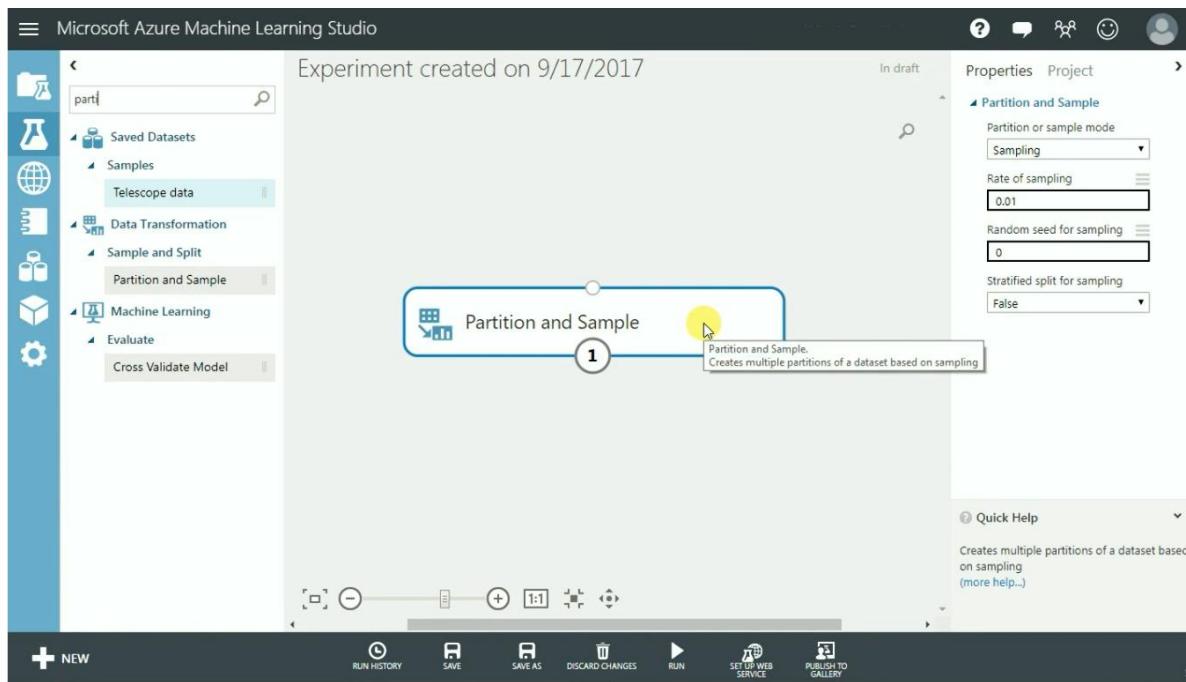


After execution visualize to check the result, Now you can see that there are no missing values in the selected columns and shows clean missing data successful



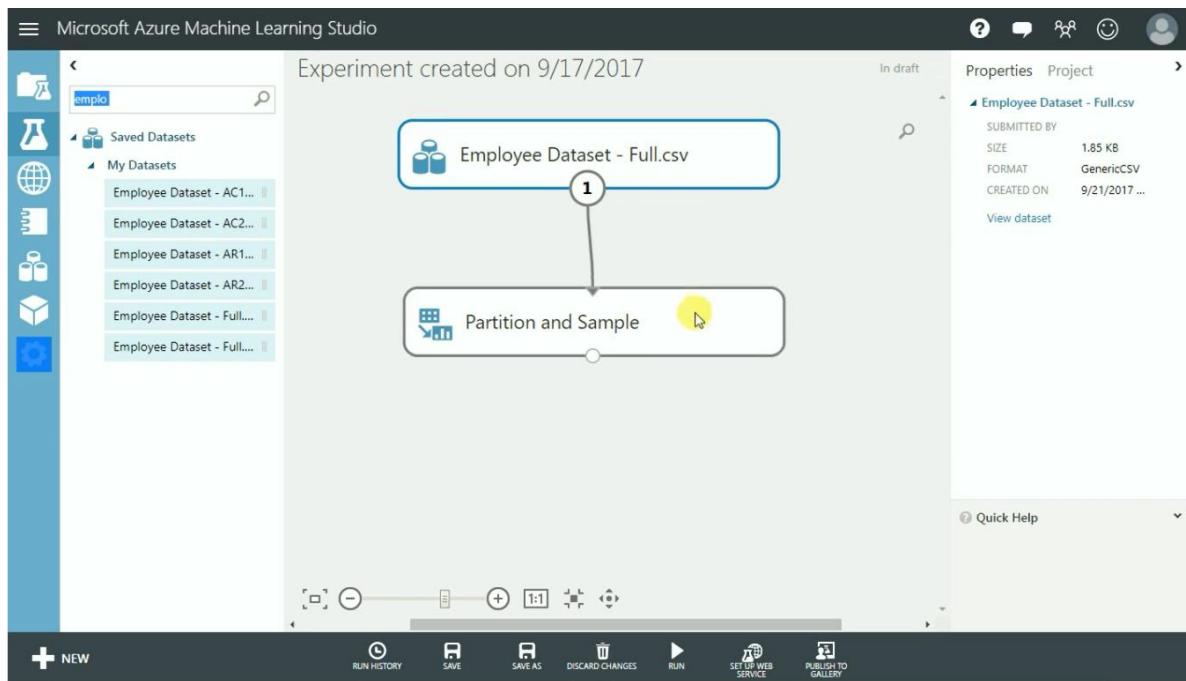
## Data Manipulation Using Partition & Sampling Component

Search for partition and sample and drop in canvas



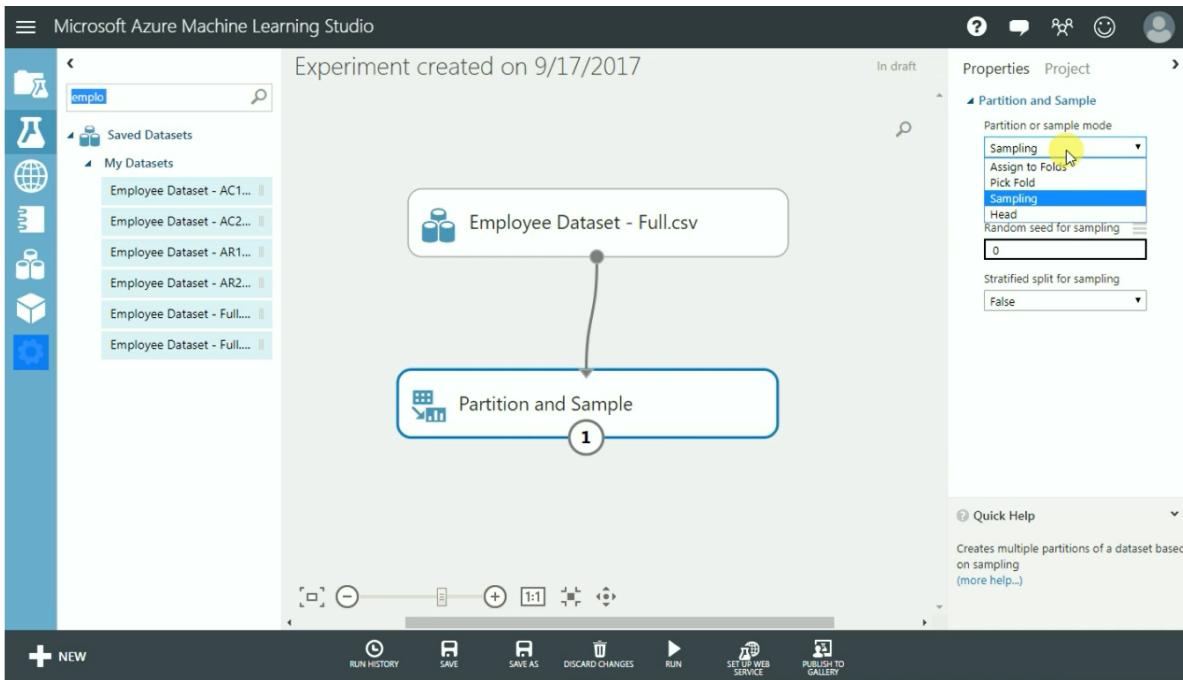
To perform partition and sample operation add an existing dataset in canvas

And connect the nodes as required

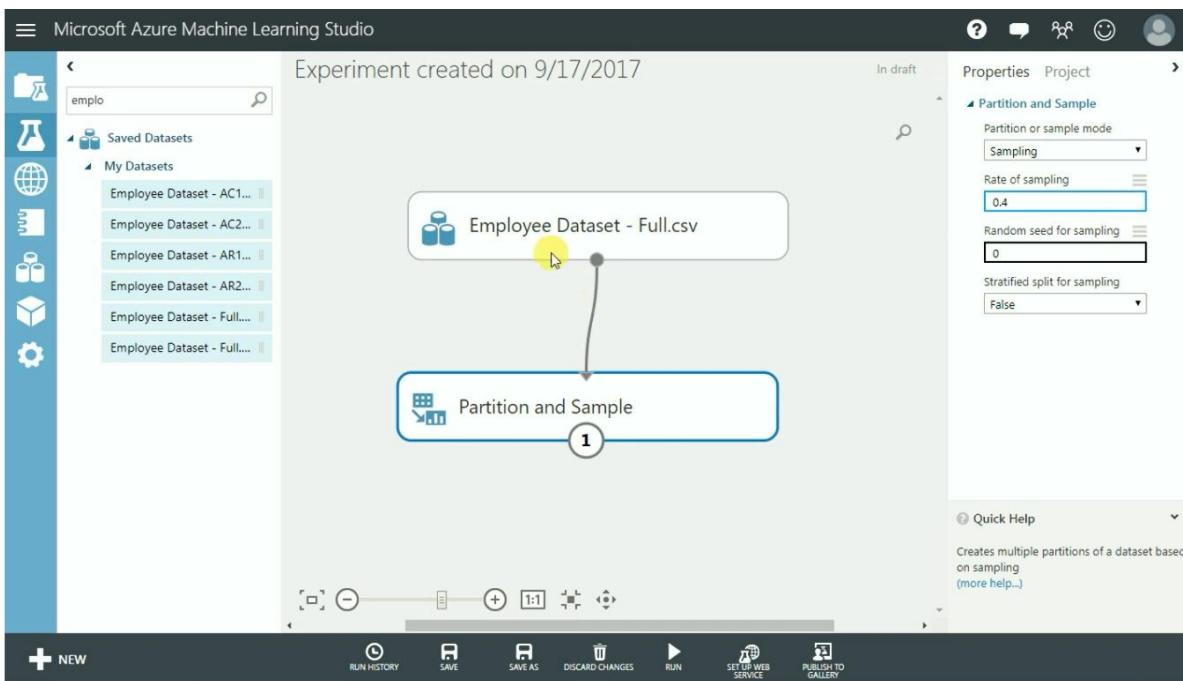


Can apply any of the types from dropdown in which we can apply

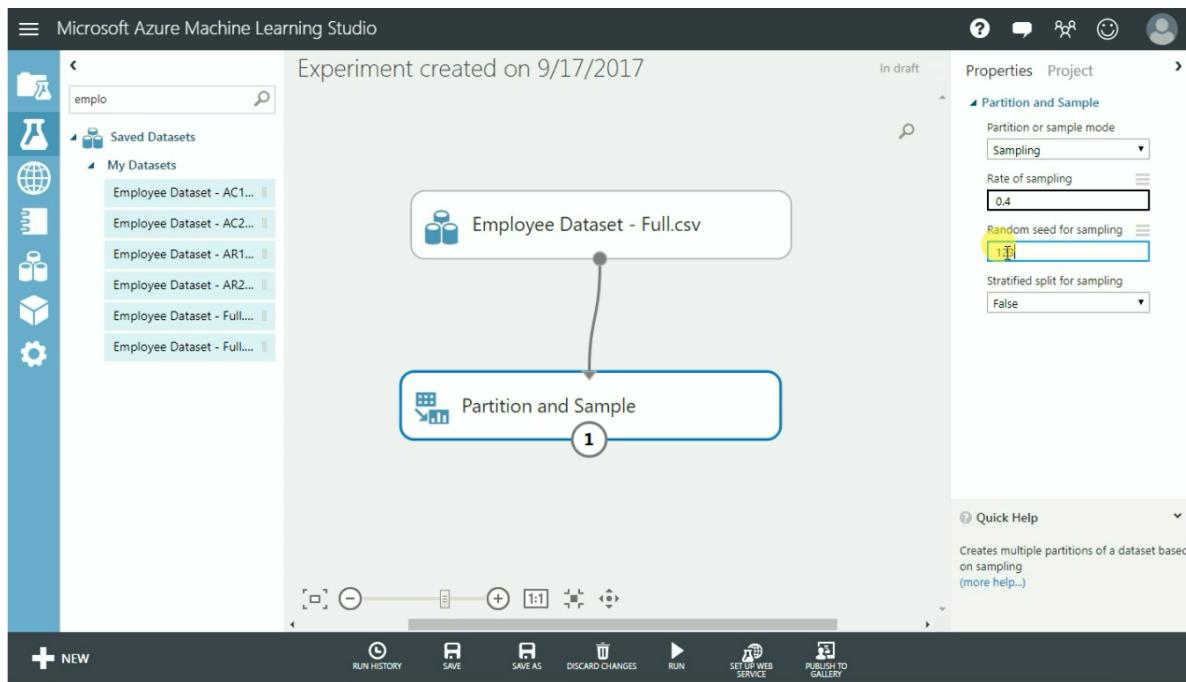
Sampling from the list



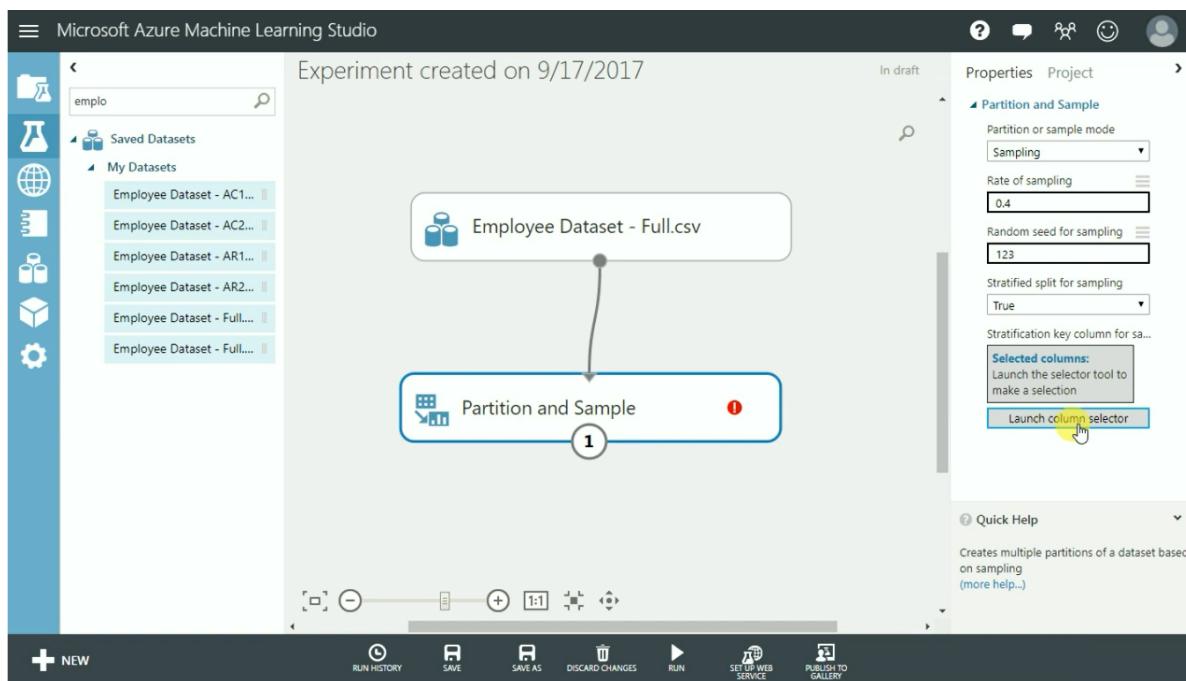
Enter the rate of sampling



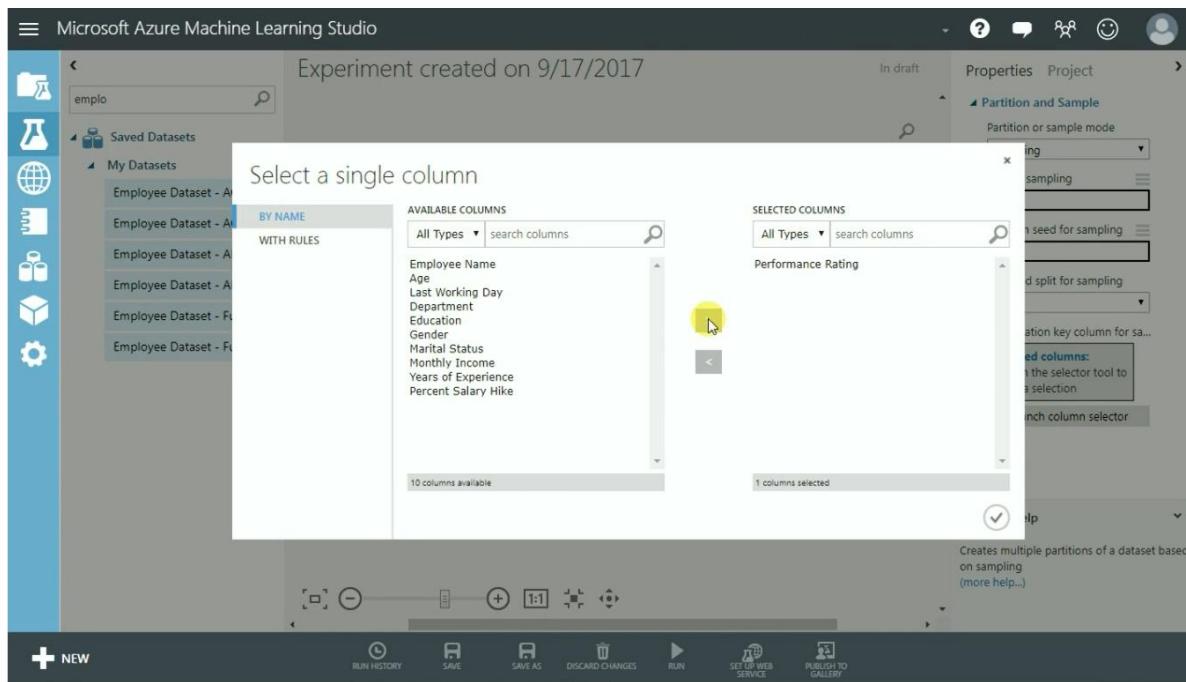
Input random seed for sampling



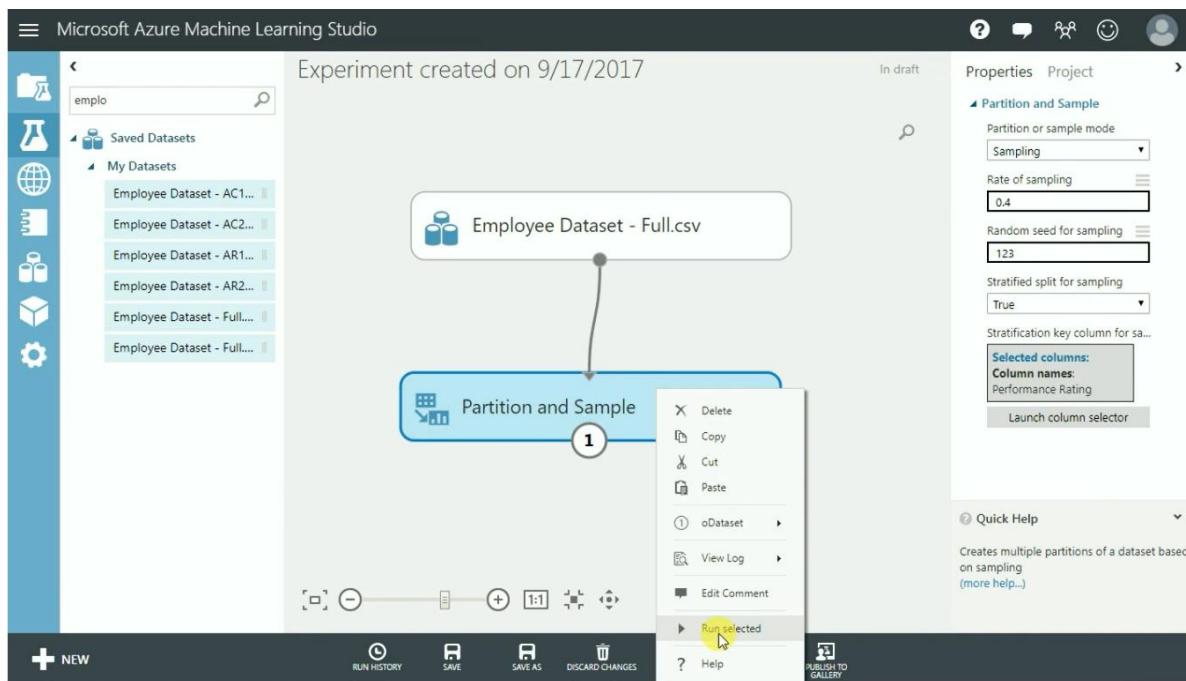
Input stratified split for sampling as true and then select launch columns selector



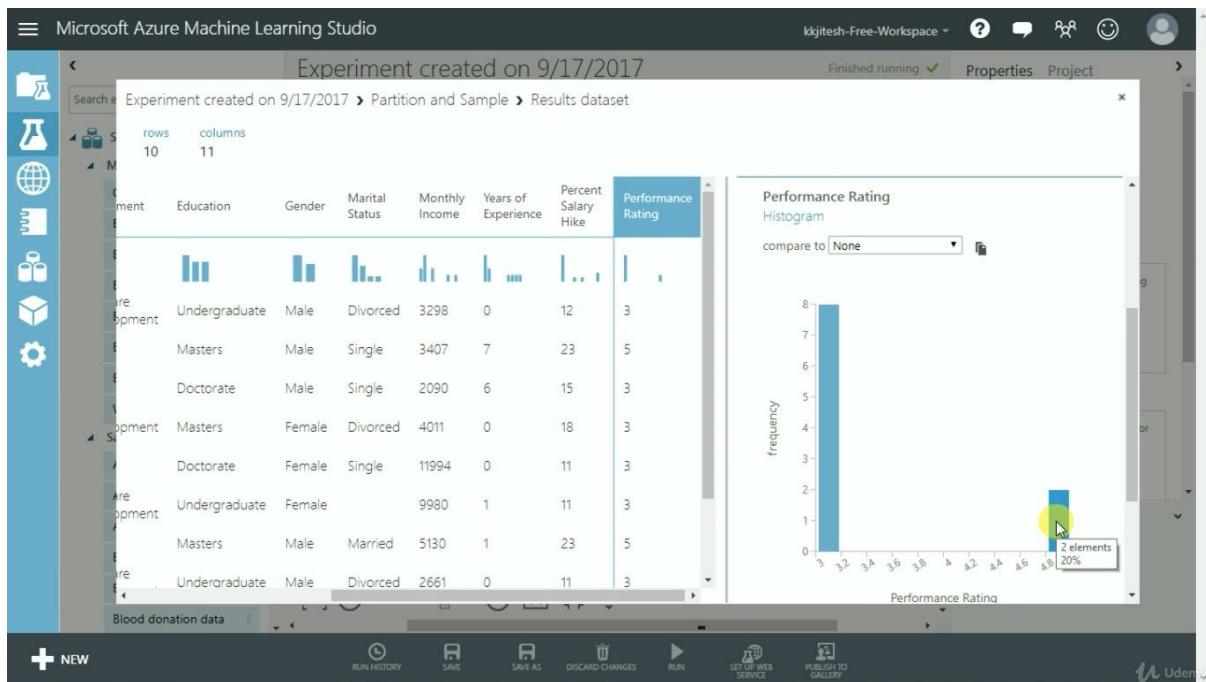
Select performance rating to the selected columns and click ok



## Run the module

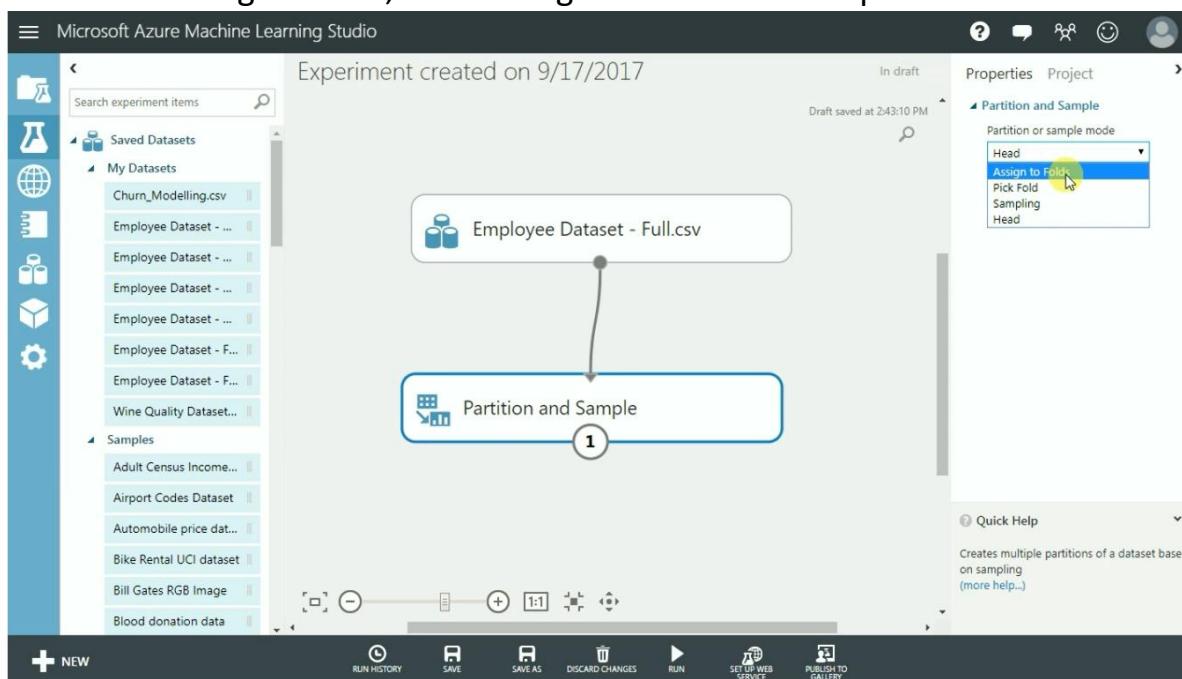


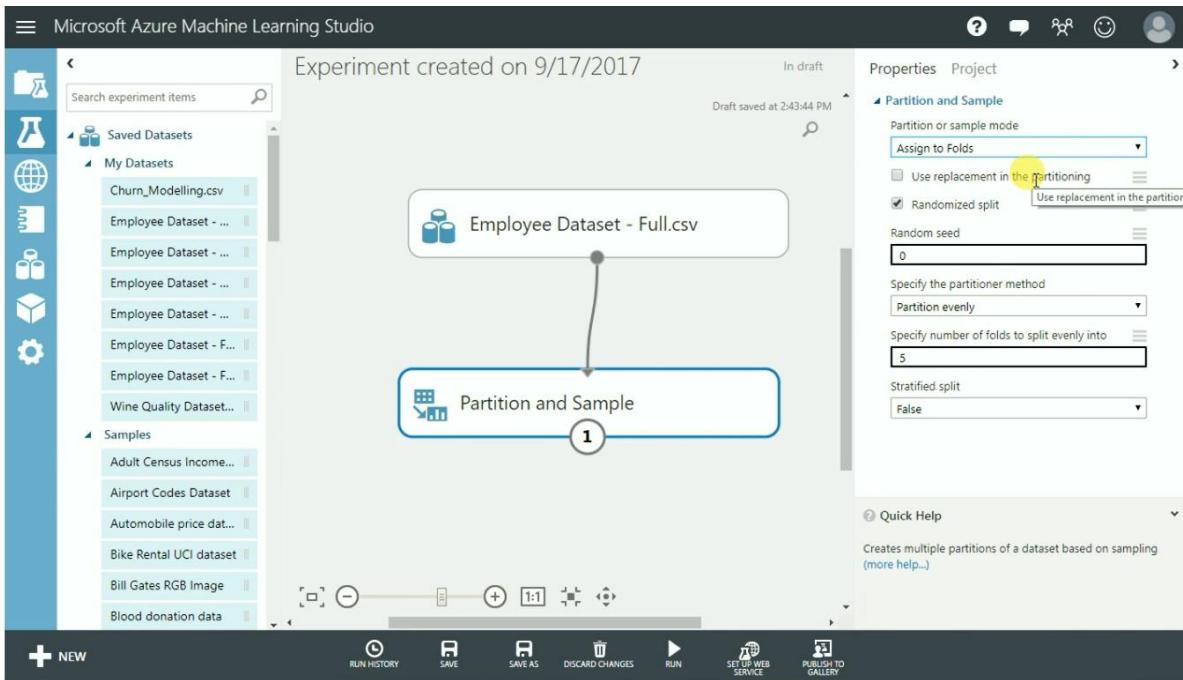
## Visualize the successful output



Another option of sample mode

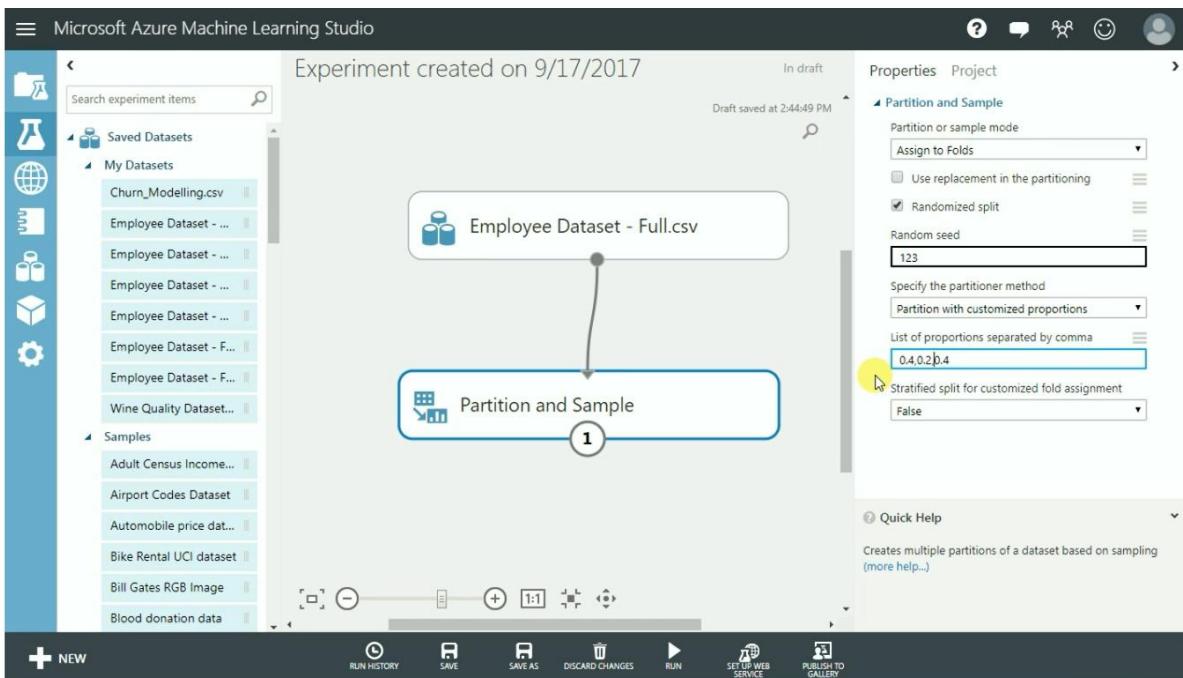
From the existing dataset , select assign to folds from dropdown



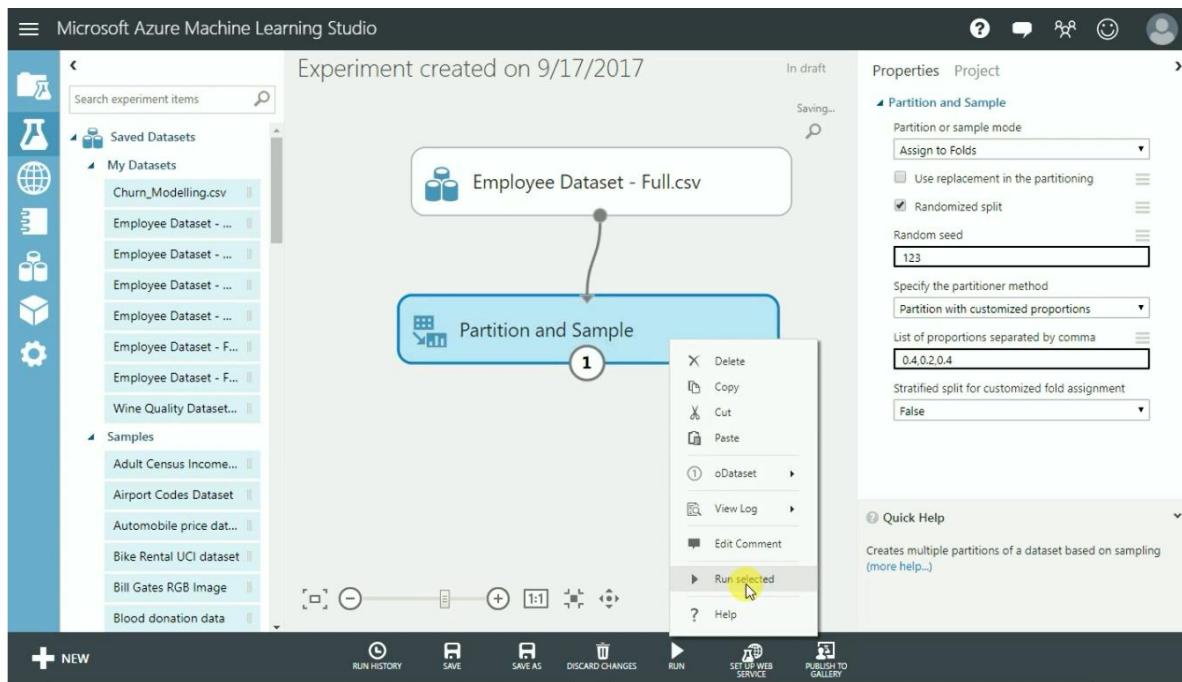


Select partition with customized proportions and list the partitions

Required separated by commas



Run the module

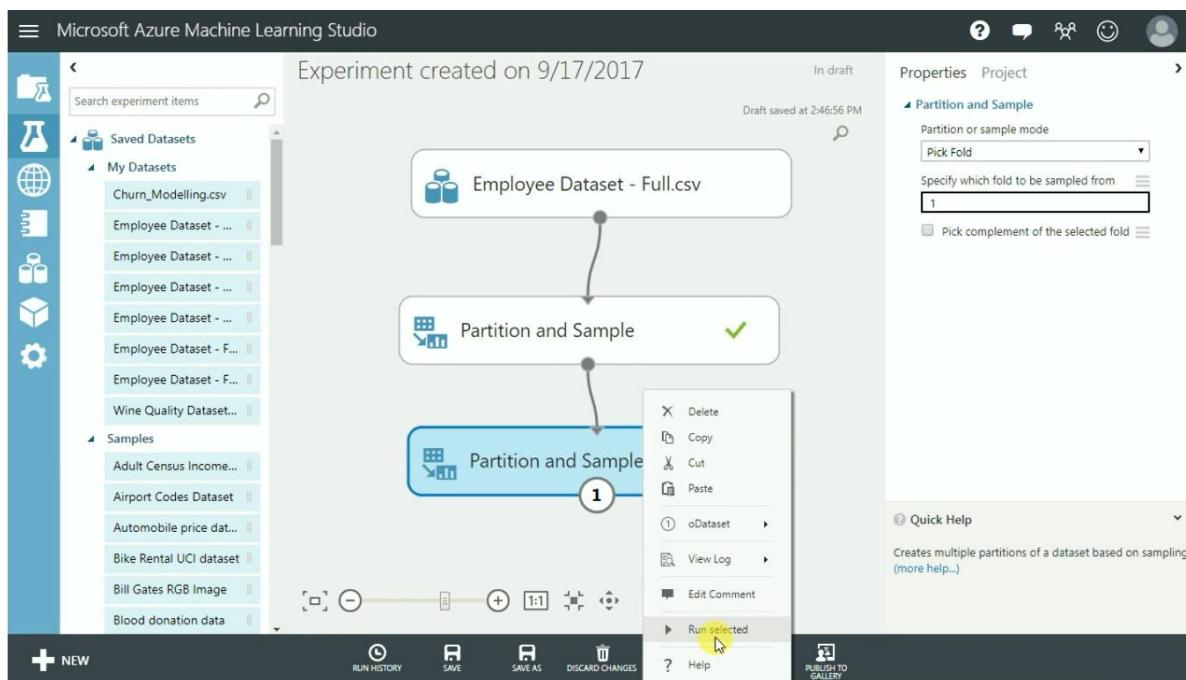


Now the visualization shows a single dataset not as three different dataset

Microsoft should provide other visualization for this module alone

Employee Name	Age	Last Working Day	Department	Education	Gender	Mar Status
Jitesh	41	31-12-9999	Training	Masters	Male	Sing
Sanjit	49	31-12-9999	Sales	Masters	Male	Mar
John	37	31-12-9999	R&D	Doctorate	Male	Sing
Sandra	33	31-12-9999	Software Development	Undergraduate	Female	Mar
Madhu	27	31-12-9999	R&D	Masters	Male	Mar
Robert	32	31-12-9999	R&D	Masters	Male	Sing
Megan	59	31-12-9999	Software Development	Masters	Female	Mar
Matt	30	31-12-2000	R&D	Doctorate	Male	Divc

Now we can try experimenting with pick fold for this copy partition and sample dataset again and select pick fold from dropdown and run the module



Now you can visualize the result after successful execution

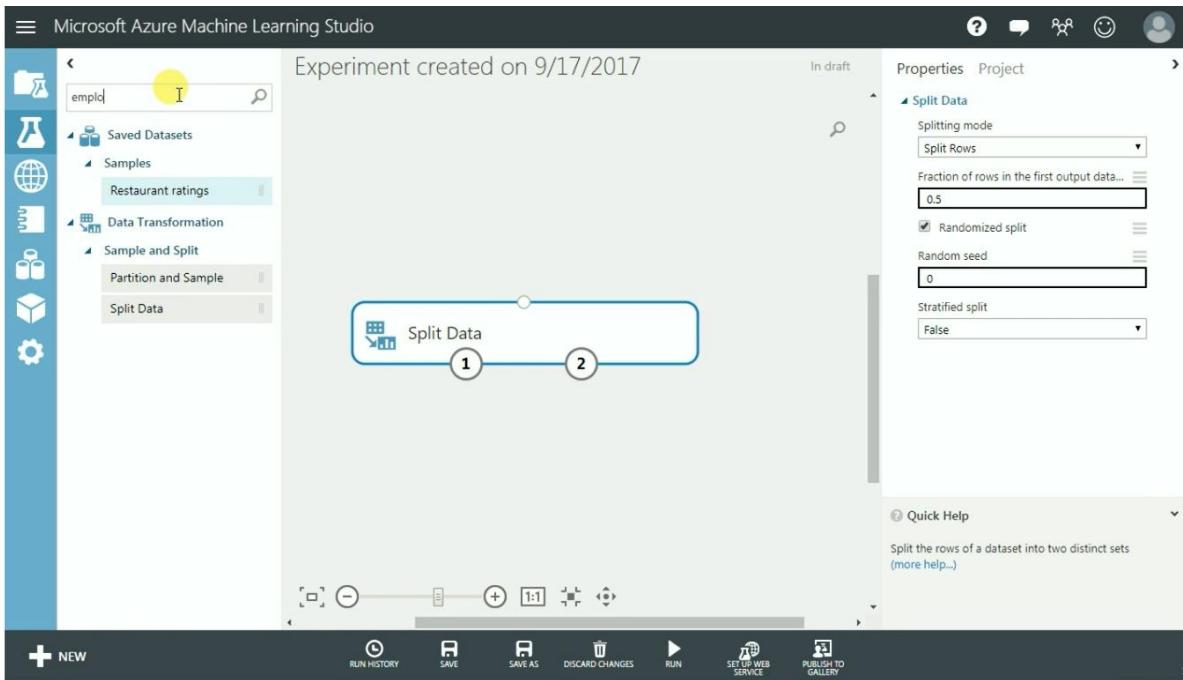
The screenshot shows the results of the experiment. The 'Partition and Sample' component is now marked as 'Finished running' with a green checkmark. The results dataset is displayed as a table with the following data:

Employee	Education	Gender	Marital Status	Monthly Income	Years of Experience	Percent Salary Hike	Performance Rating
1	Masters	Male	Divorced	2935	1	13	3
2	Masters	Male	Married	2426	0	13	3
3	Doctorate	Male	Single	2090	6	15	3
4	Masters	Female	Divorced	4011	0	18	3
5	Doctorate	Female	Single	11994	0	11	3
6	Undergraduate	Female	Divorced	9980	1	11	3
7	Undergraduate	Female	Married	2909	1	11	3
8	Undergraduate	Male	Divorced	2661	0	11	3

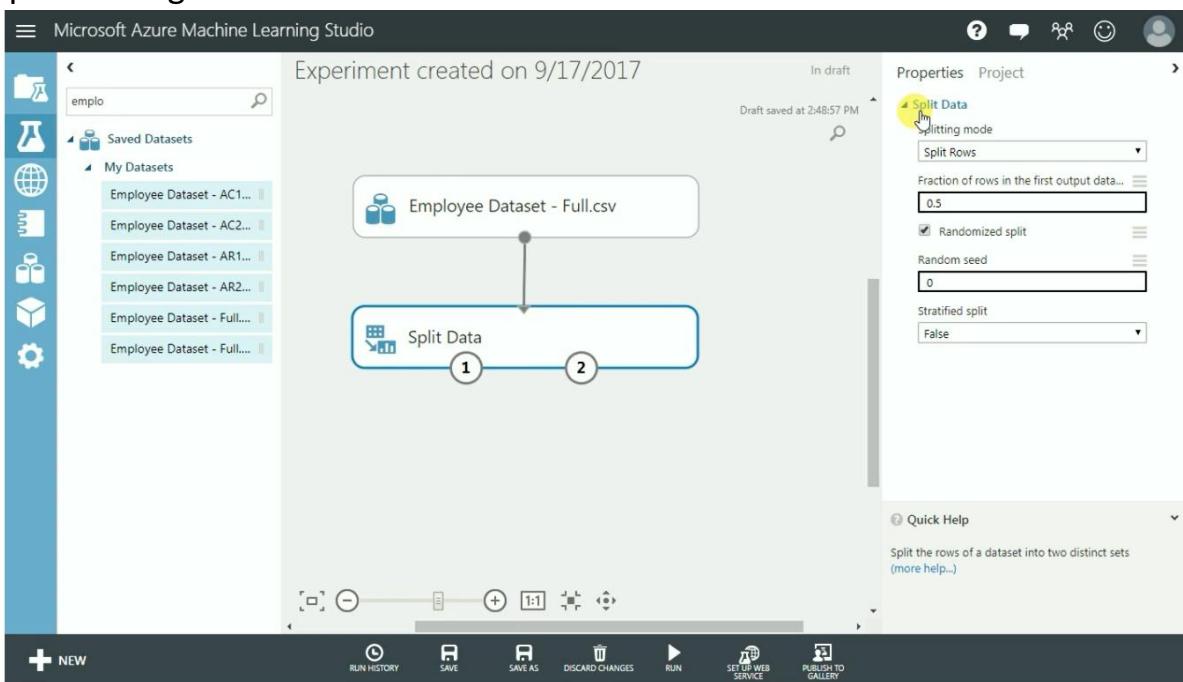
To the right, there are sections for 'Statistics' and 'Visualizations'. The 'Visualizations' section includes a note: 'To view, select a column in the table.'

## Data Manipulation Using Split Data Component

Search for split data and drop it in the canvas

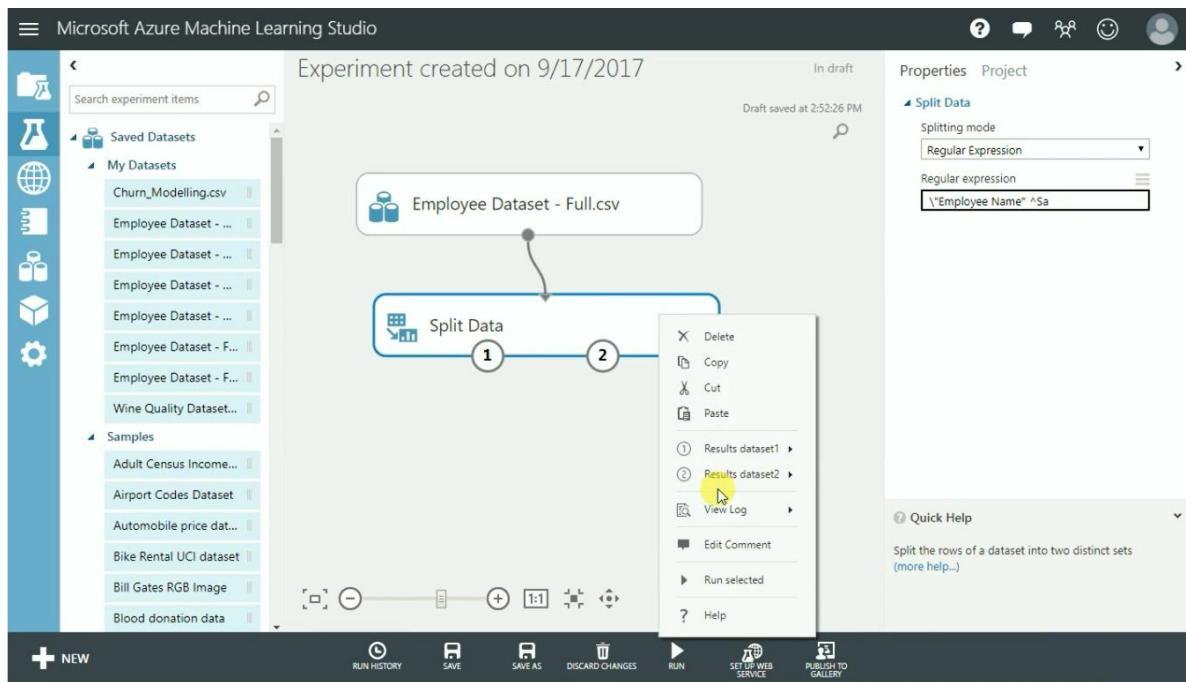


**Input existing dataset and connect the nodes**

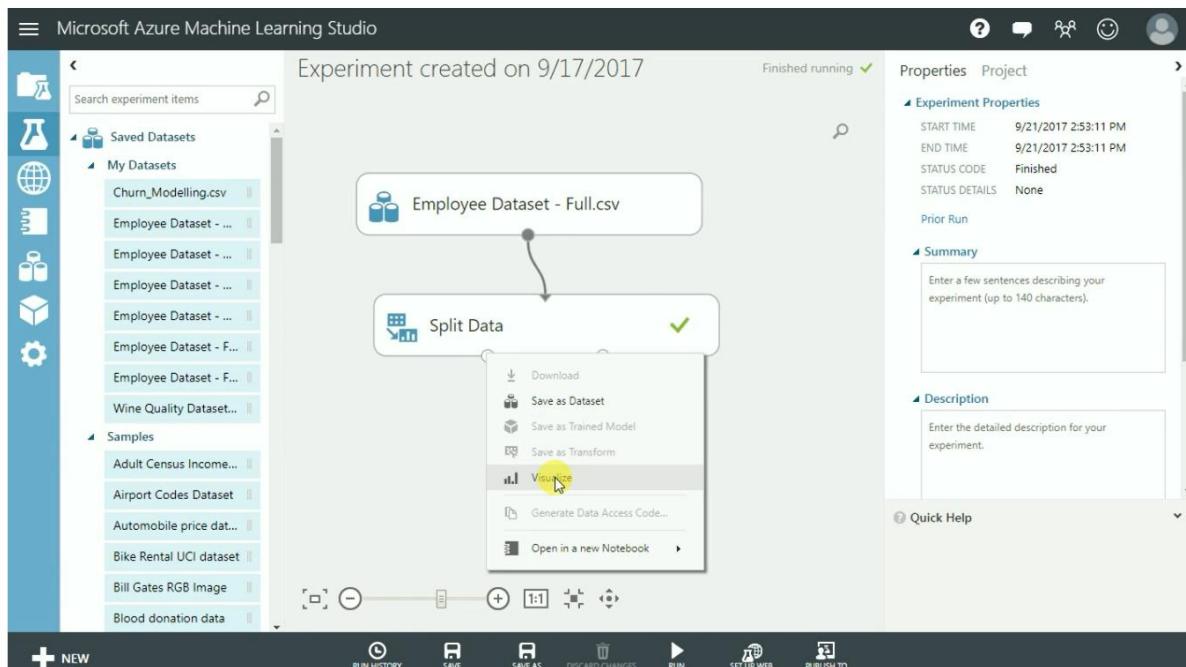


**Select parameters as regular expression from drop down and enter**

**Regular expression with name starting with 'sa' and run the module**



Visualize the output after successful run



Result successful

Experiment created on 9/17/2017 > Split Data > Results dataset1

rows 2 columns 11

Employee Name	Age	Last Working Day	Department	Education	Gender	Marital Status
Sanjiti	49	31-12-9999	Sales	Masters	Male	Married
Sandra	33	31-12-9999	Software Development	Undergraduate	Female	Married

Similarly try splitting with relative expression example for employee's age

Greater than 30

Experiment created on 9/17/2017

In draft

Draft saved at 2:54:45 PM

Properties Project

Split Data

Splitting mode: Relative Expression

Relational expression: \\"Age" > 30

```

graph TD
    A[Employee Dataset - Full.csv] --> B[Split Data]
    B -- 1 --> C
    B -- 2 --> D
  
```

Run the module and visualize the output in node1 for age more than 30

Microsoft Azure Machine Learning Studio

Experiment created on 9/17/2017

rows 18 columns 11

Employee Name	Age	Last Working Day	Department	Education	Gender	Marital Status
Jitesh	41	31-12-9999	Training	Masters	Male	Sing
Sanjit	49	31-12-9999	Sales	Masters	Male	Mar
John	37	31-12-9999	R&D	Doctorate	Male	Sing
Sandra	33	31-12-9999	Software Development	Undergraduate	Female	Mar
Robert	32	31-12-9999	R&D	Masters	Male	Sing
Megan	30	31-12-9999	Software Development	Masters	Female	Mar
Will	38	01-03-2012	R&D	Doctorate	Male	Sing
George	36	31-12-9999	Software Development	Masters	Male	Mar

Blood donation data

view as

To view, select a column in the table.

Statistics

Visualizations

Run History Save Save As Discard Changes Run Set Up Web Service Publish To Gallery

Check node 2 for rest of the dataset

Microsoft Azure Machine Learning Studio

Experiment created on 9/17/2017

rows 7 columns 11

Employee Name	Age	Last Working Day	Department	Education	Gender	Marital Status
Madhu	27	31-12-9999	R&D	Masters	Male	Married
Matt	30	31-12-9999	R&D	Doctorate	Male	Divorced
Emma	29	31-12-9999	Software Development	Undergraduate	Female	Single
Clint	28	31-12-9999	Software Development	Undergraduate	Male	Single
Kate	29	31-12-9999	Software Development	Undergraduate	Female	Single
Mel	22	01-06-2014	Software Development	Masters	Male	Divorced
Katherine	24	31-12-9999	Development	Masters	Female	Divorced

Blood donation data

view as

To view, select a column in the table.

Statistics

Visualizations

Run History Save Save As Discard Changes Run Set Up Web Service Publish To Gallery