Microsoft

Data Processing with Azure ML

Table of Contents

Goals and Requirements

Estimated time to complete lab is 40-45 minutes.

Goals

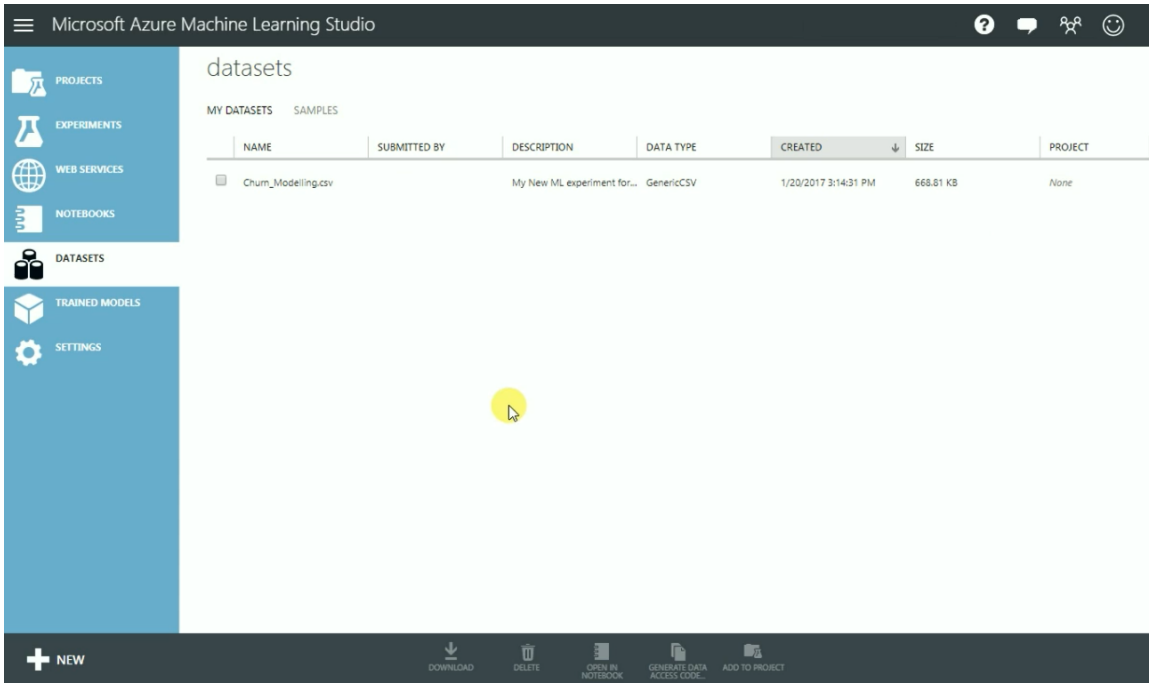
1. Introduce to Azure Machine Learning Environment
2. Azure Machine Learning Data Pre-processing and Cleansing.
3. Implementing Data Transformations.

Requirements

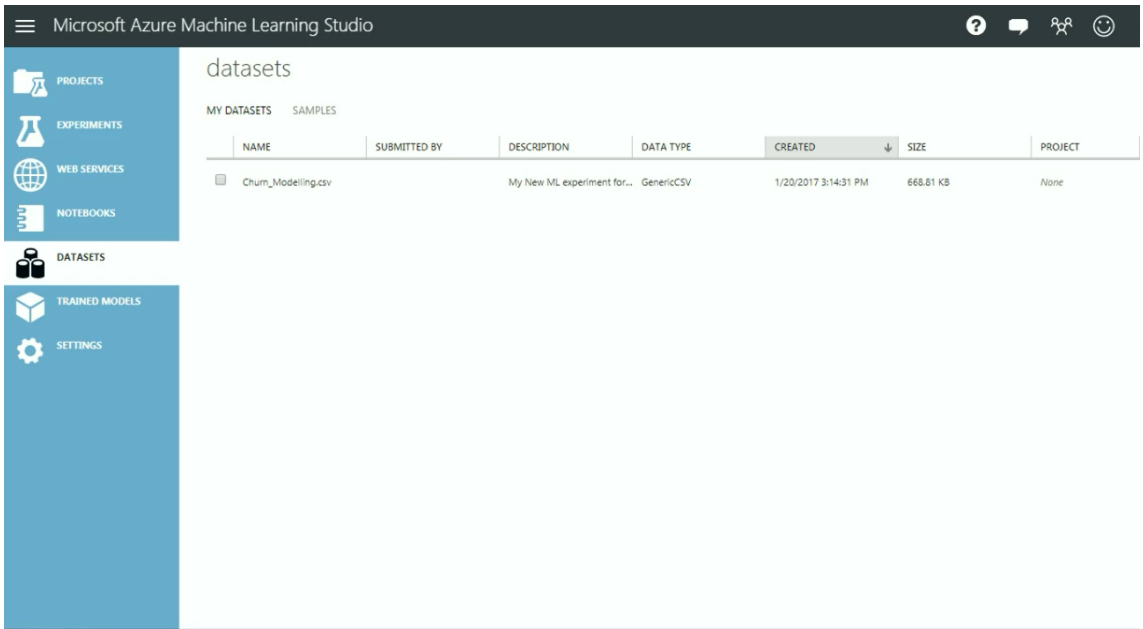
1. Access to an Azure Machine Learning Studio and related documents provided in the session.

Data Input Output to Azure Workspace

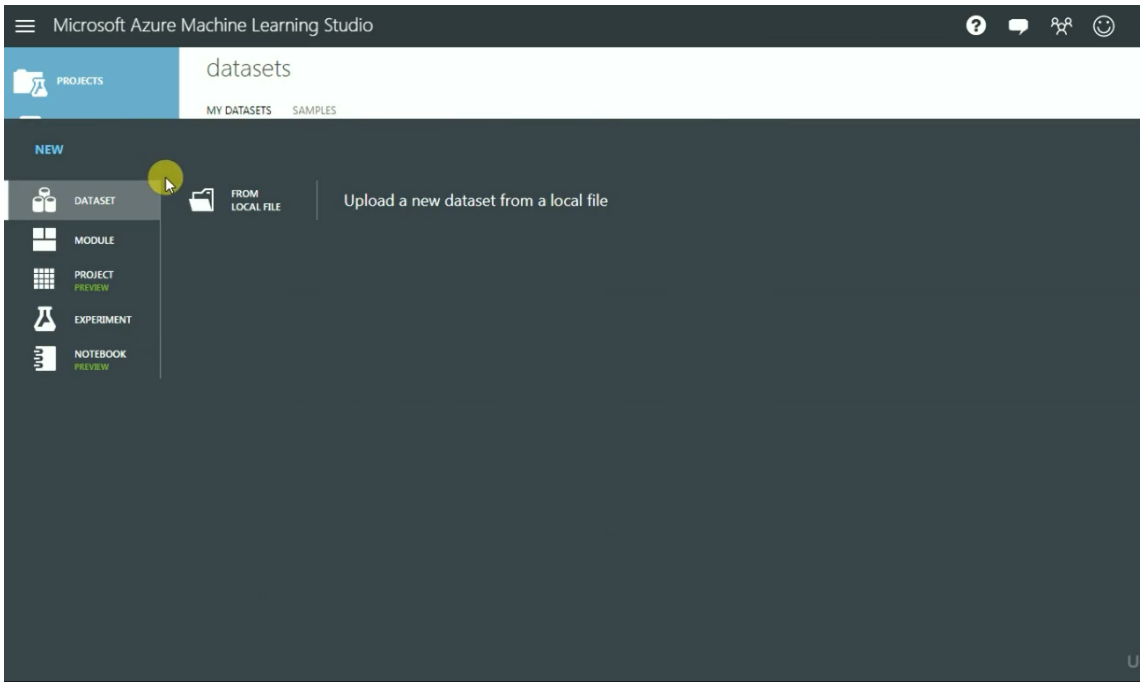
Data Importing Process



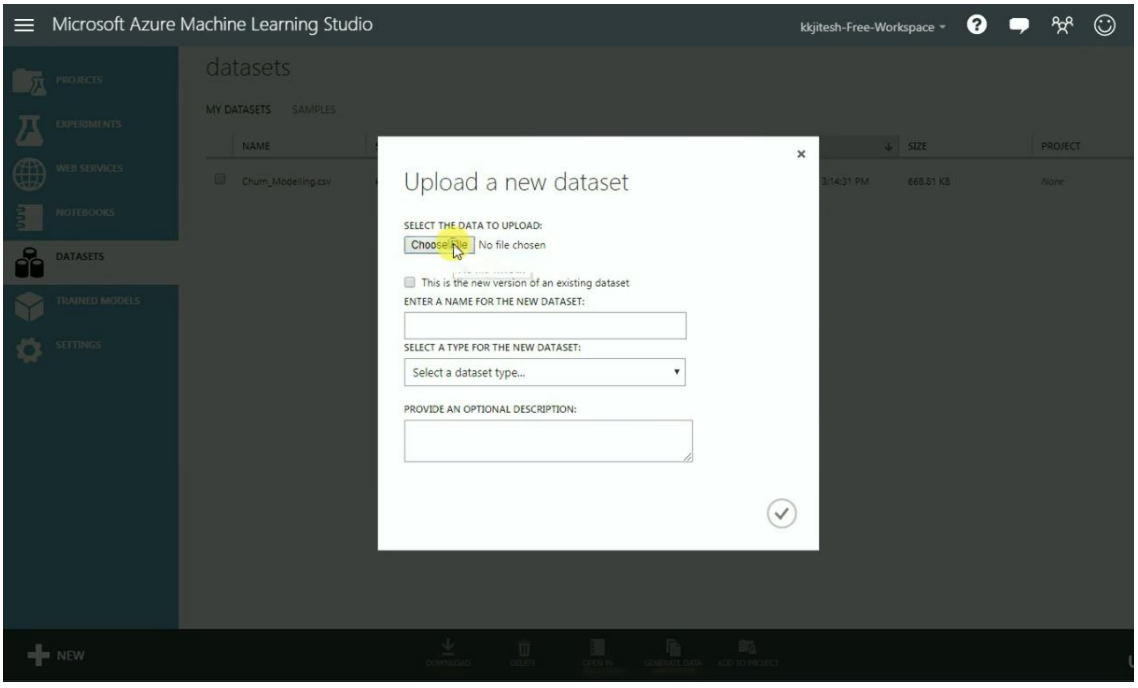
Click on New



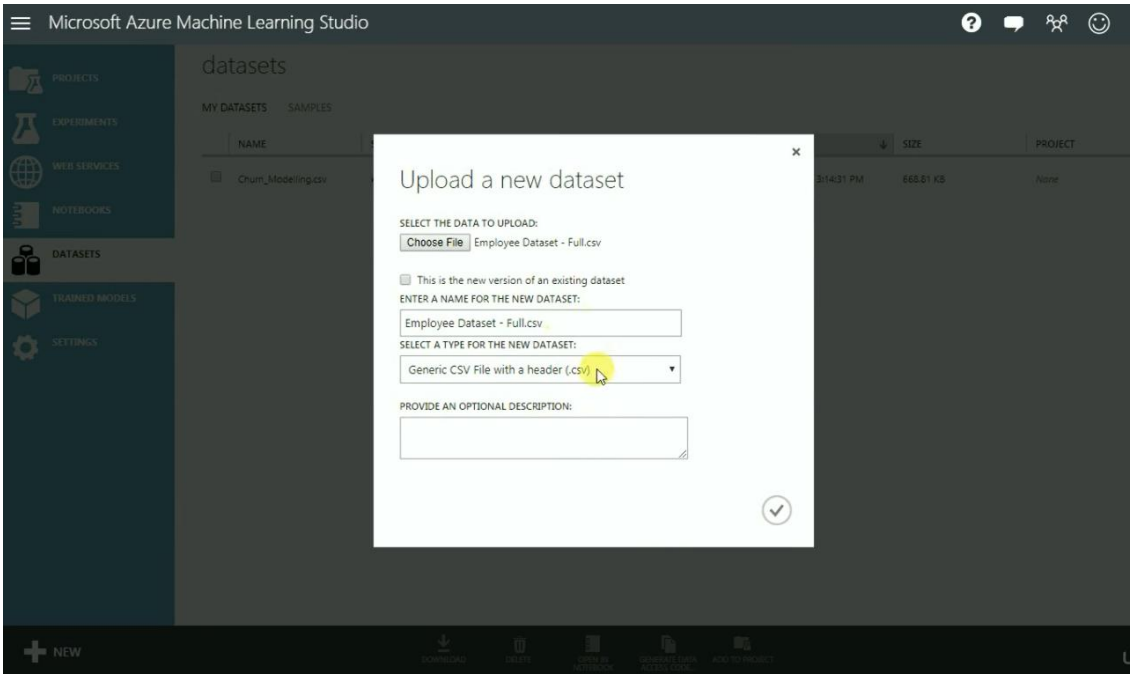
Go to dataset and click on from Local File



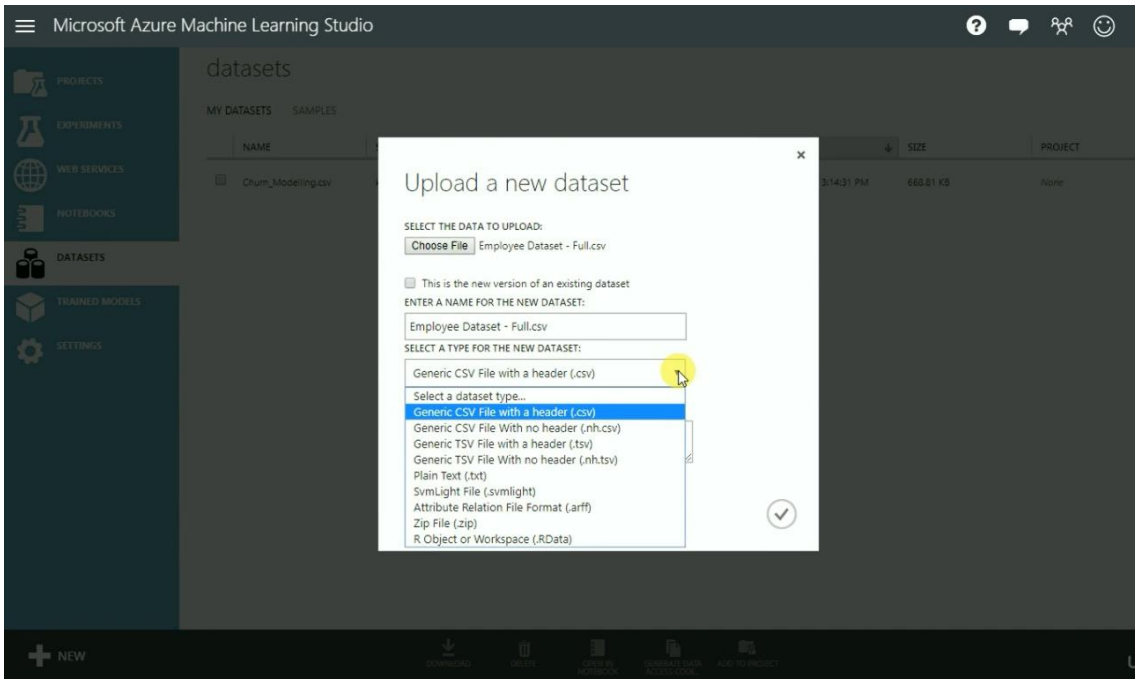
Click on choose a file



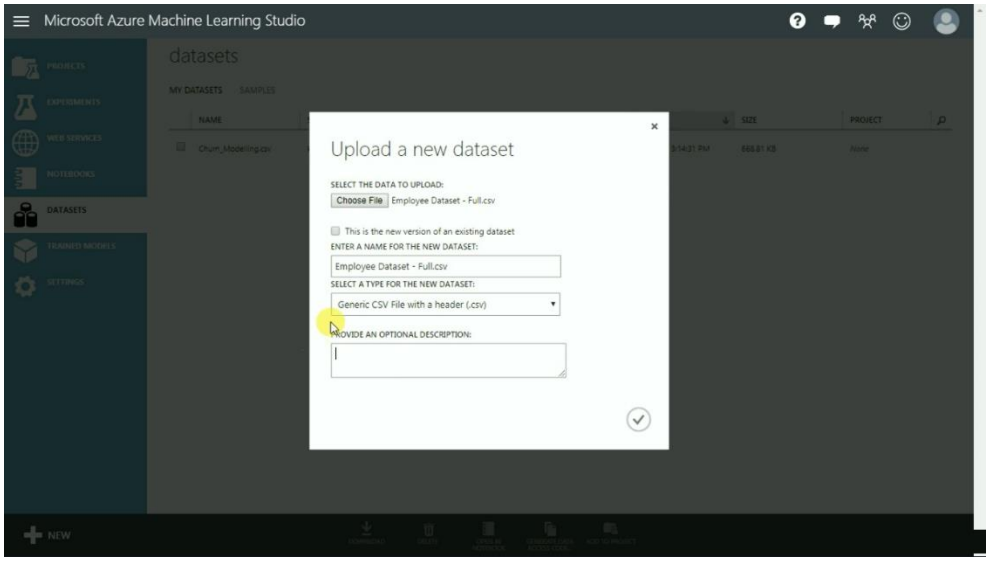
Select the dataset Employee-Dataset-Full.csv from the local folder shared.



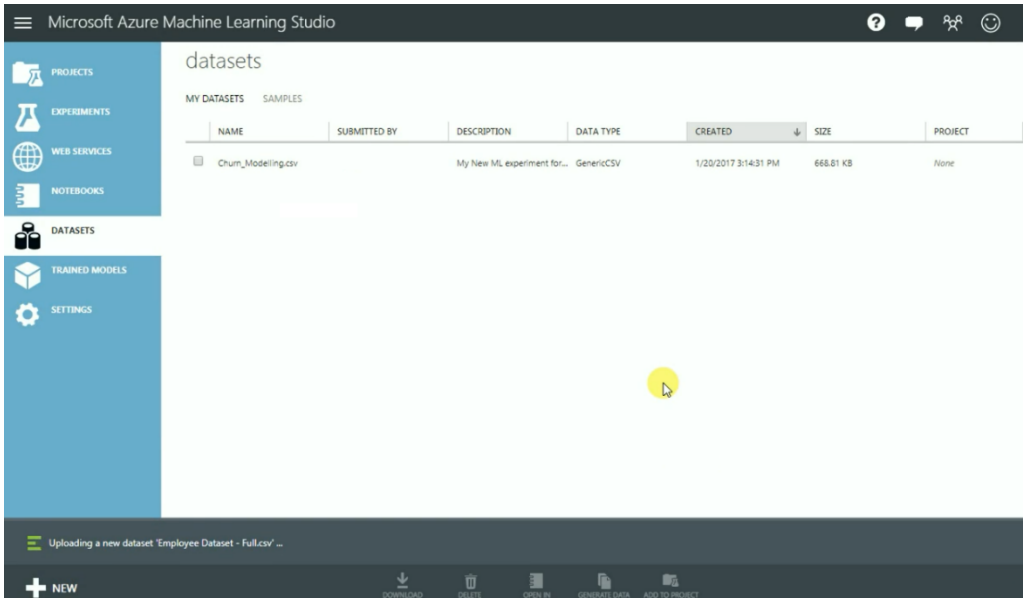
Can select any options from the dropdown menu



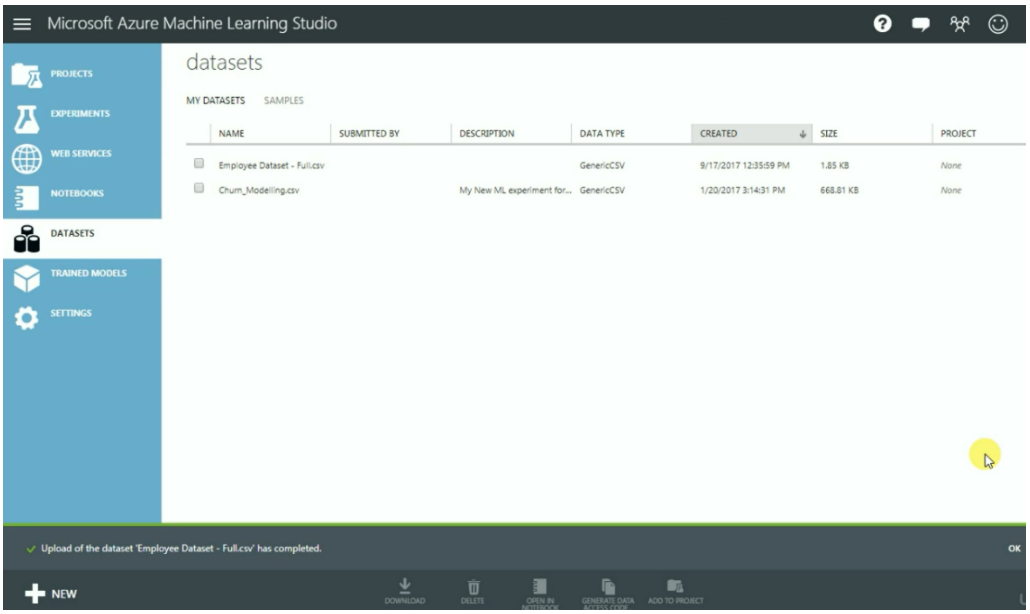
Can provide additional description also as shown below and then click ok



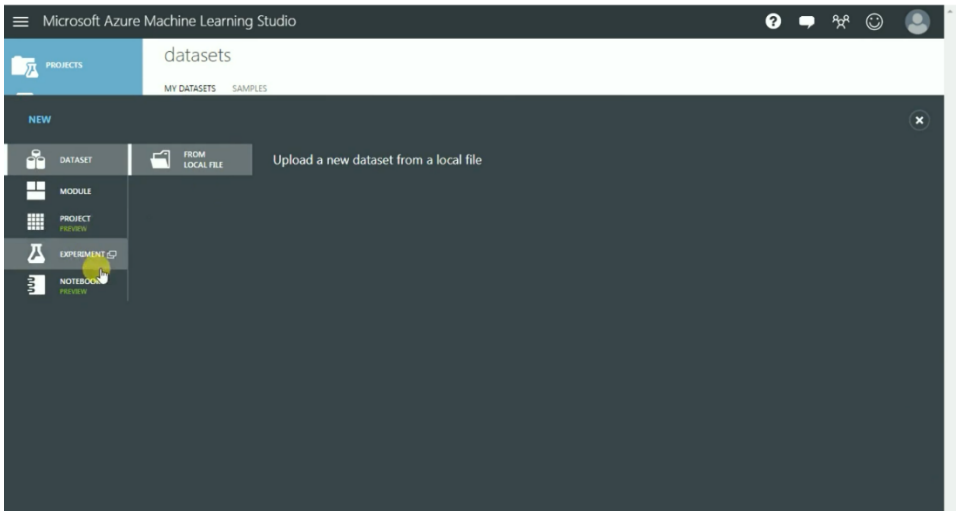
Wait for dataset update



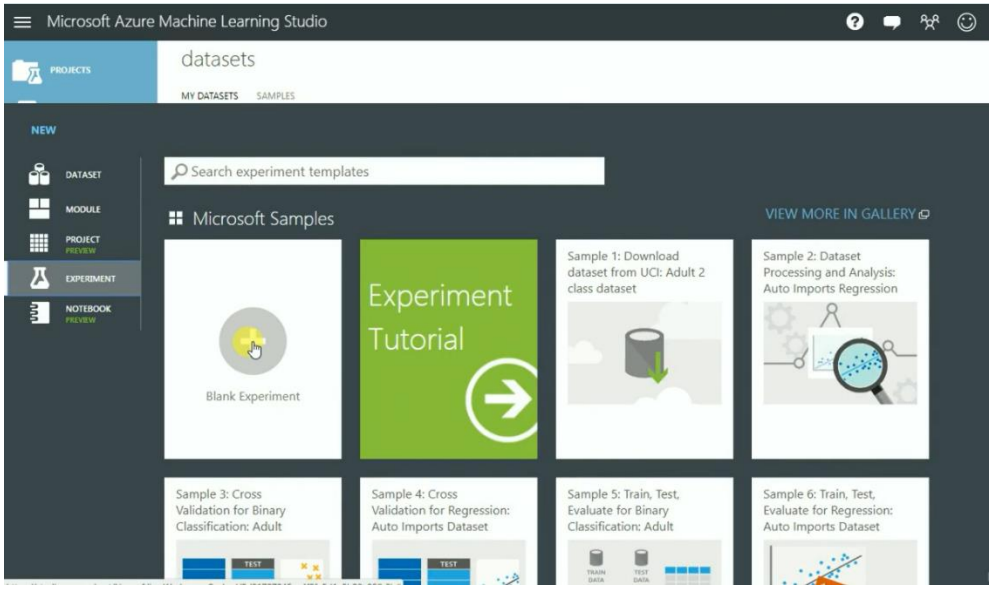
After updating click ok and then click new



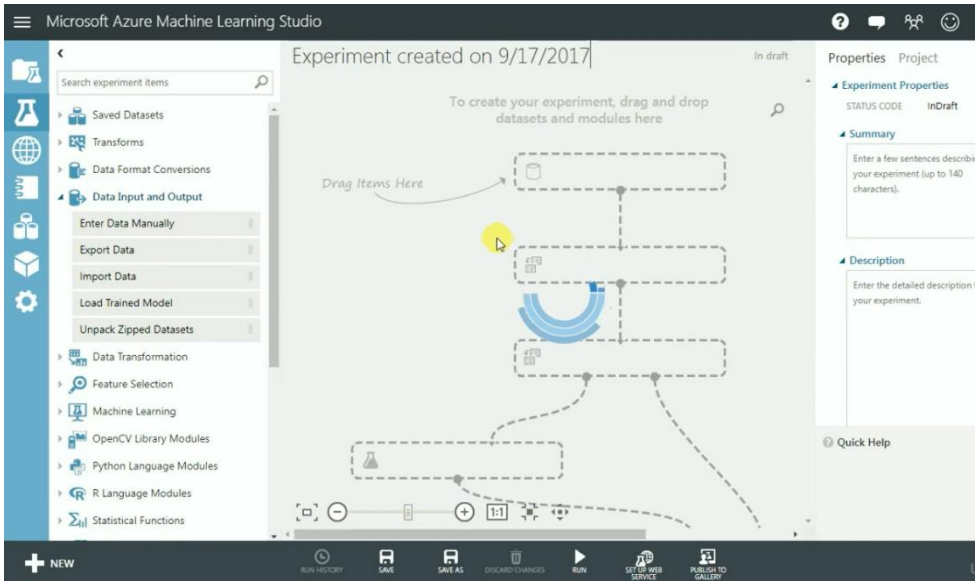
Creation of Experiment



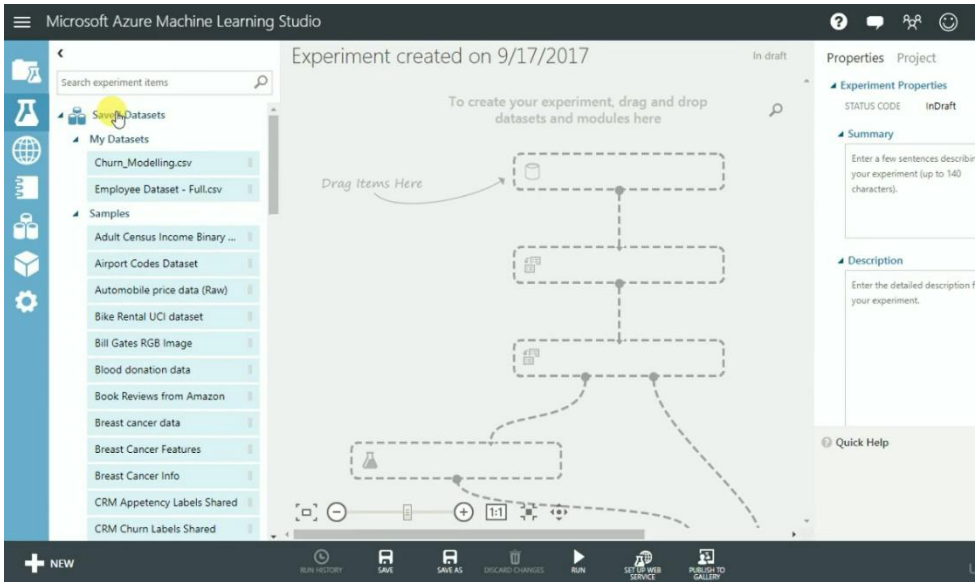
Click experiment and blank experiment



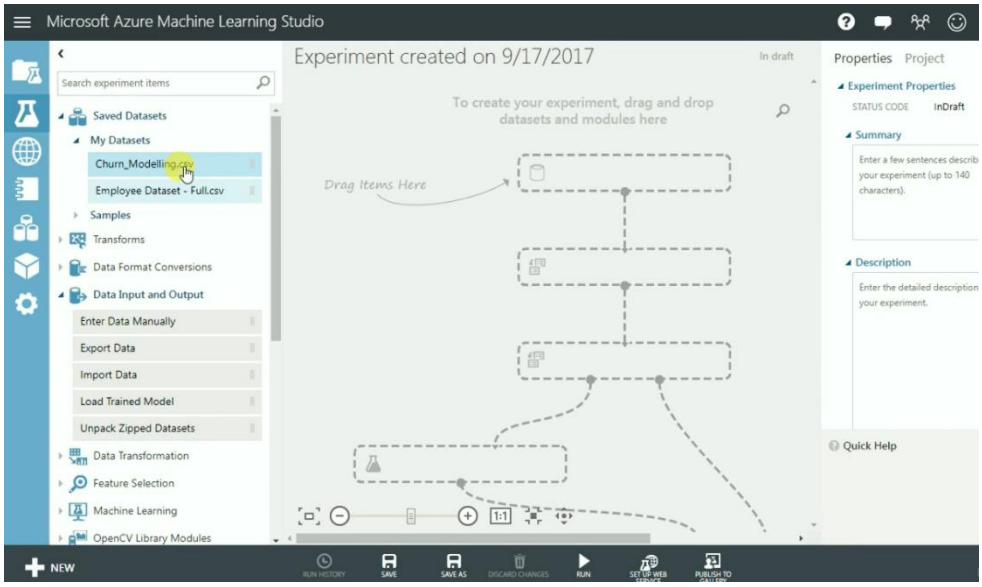
Check the datasets



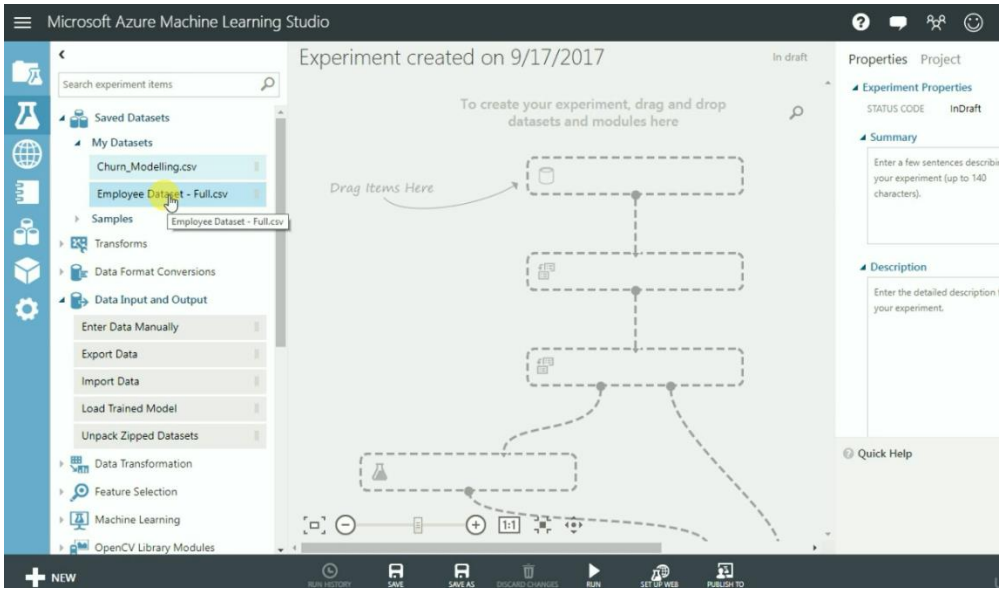
Click on saved data sets



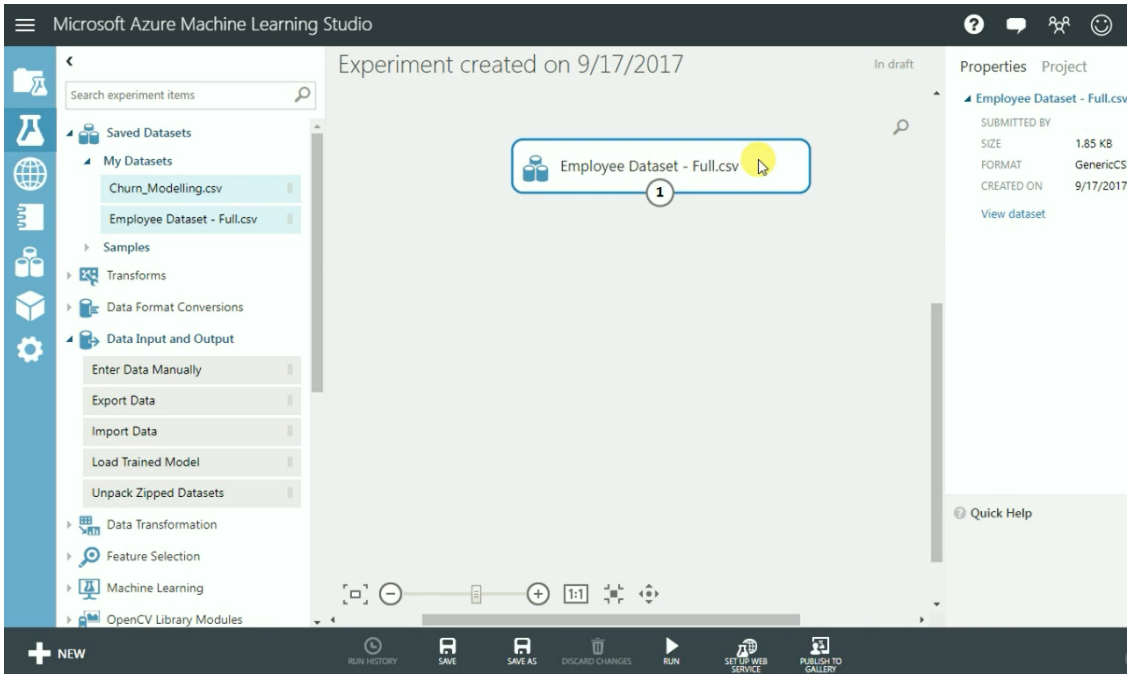
Dataset updated



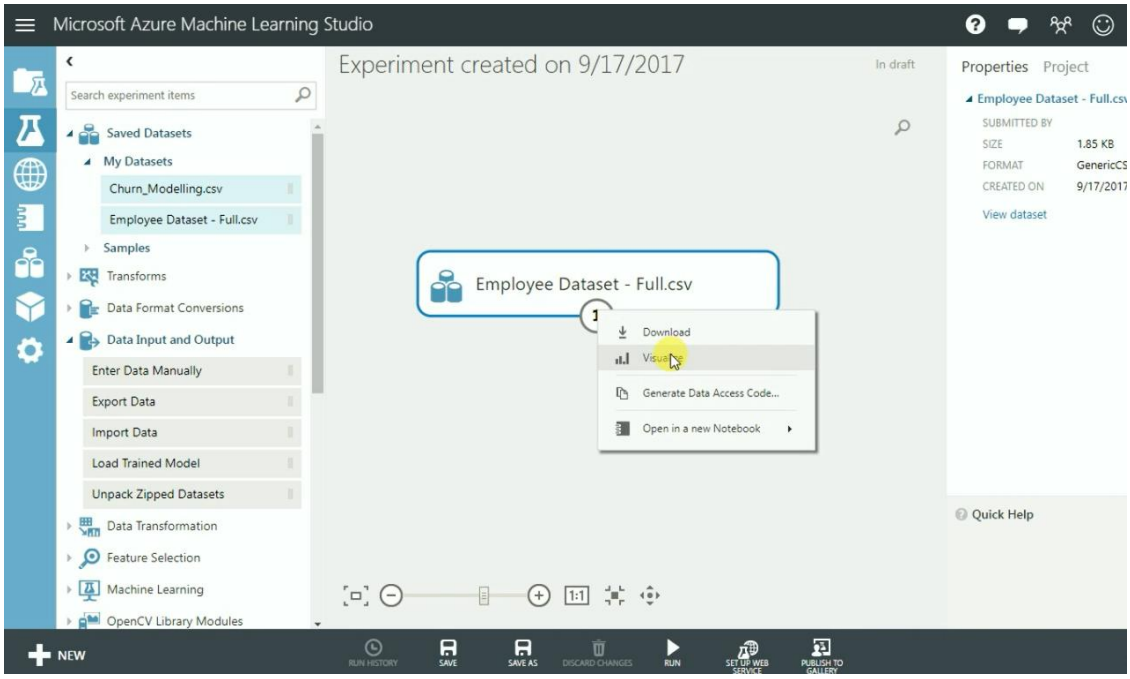
Post updating drag and drop the data set

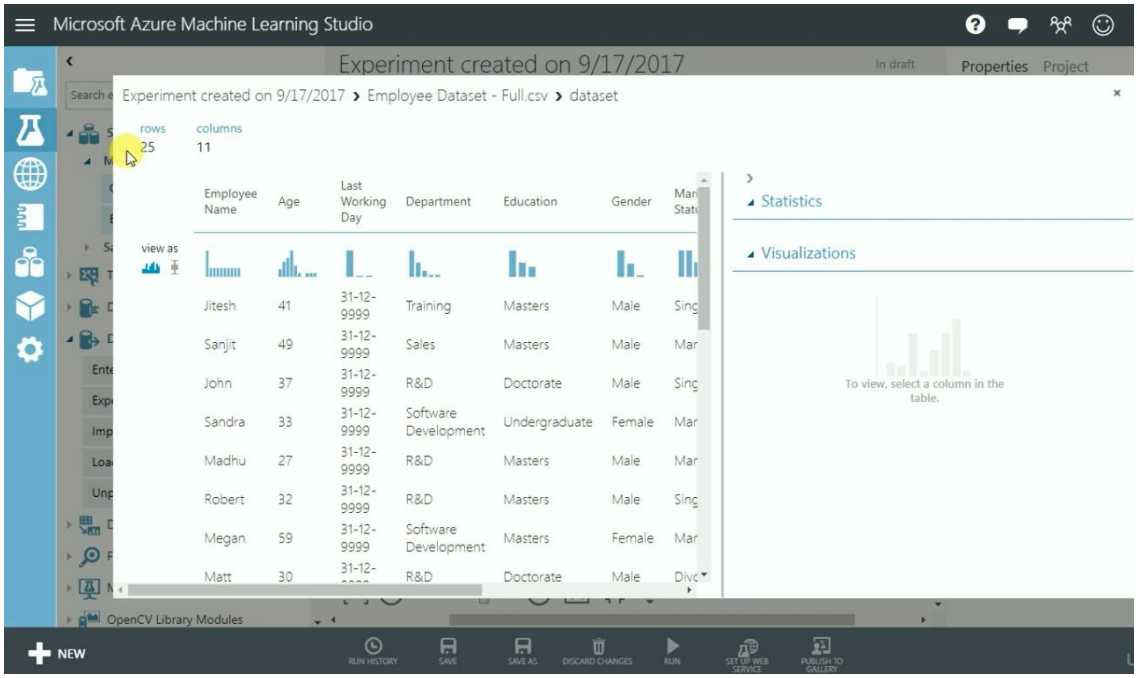


You can find flowchart structured rectangle with a node

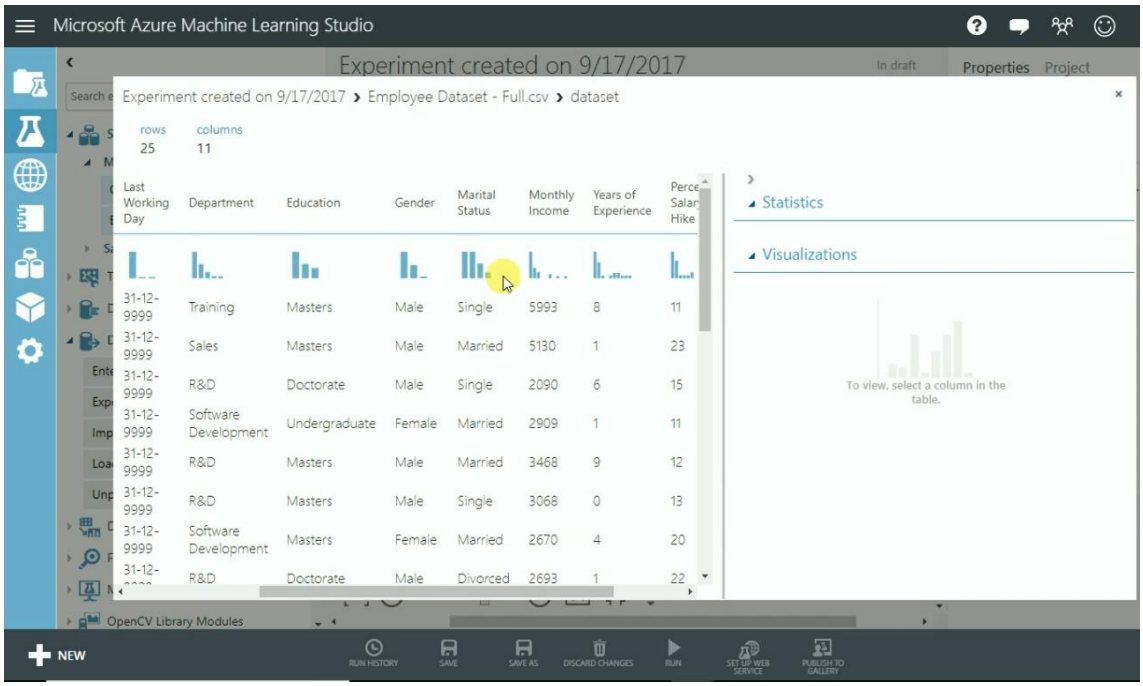


Right Click on node 1 and then click visualize

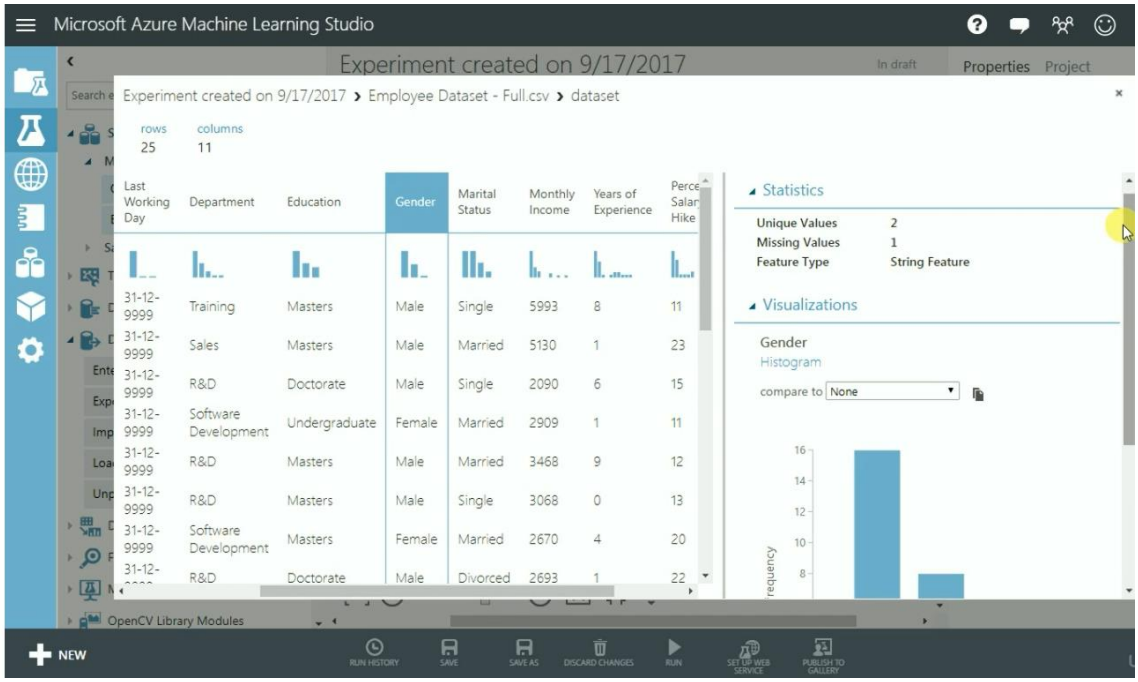




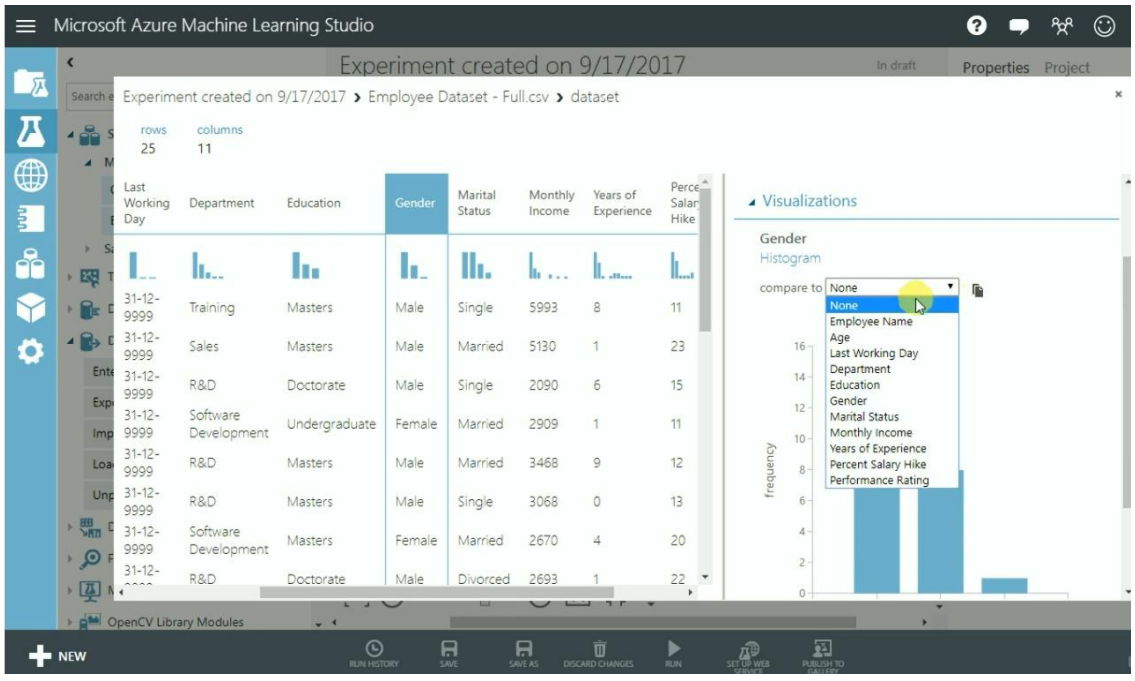
Can also view the graph



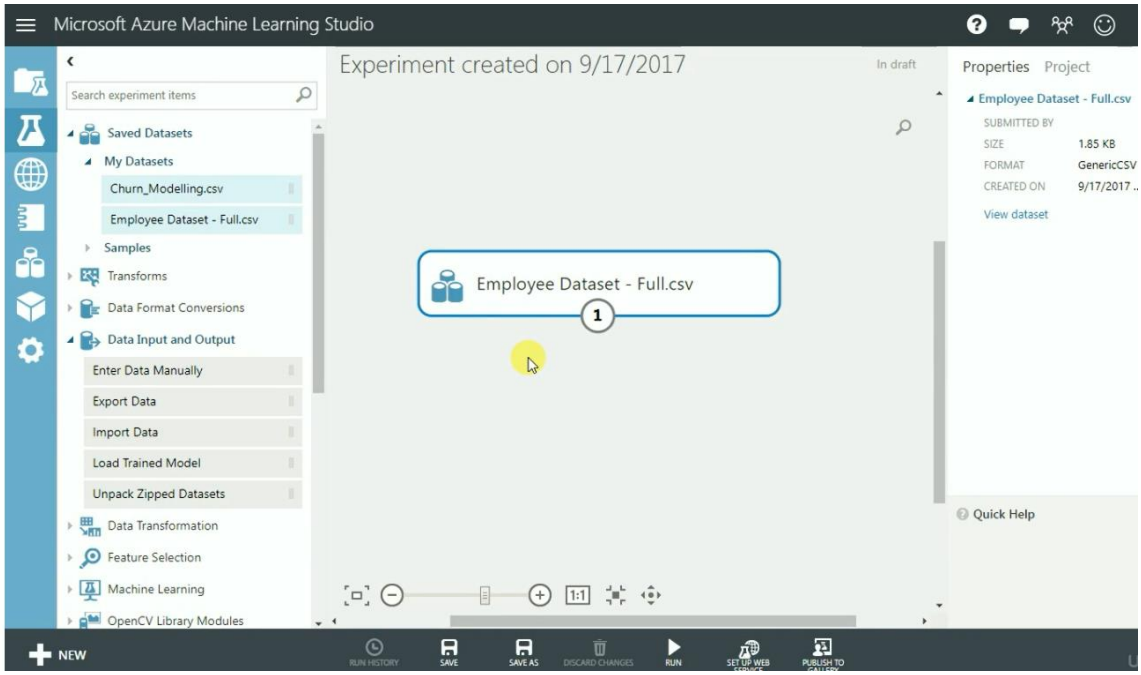
You can view statistics and graph if any data selected



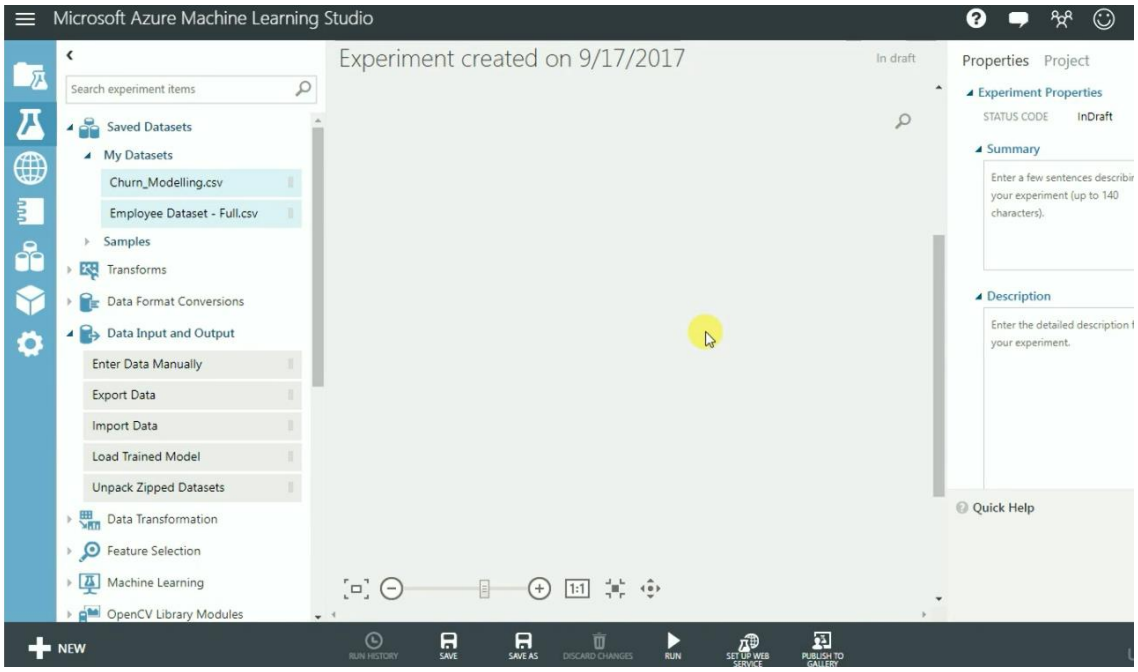
You can compare the data from dropdown



Close the visualization and go back to input

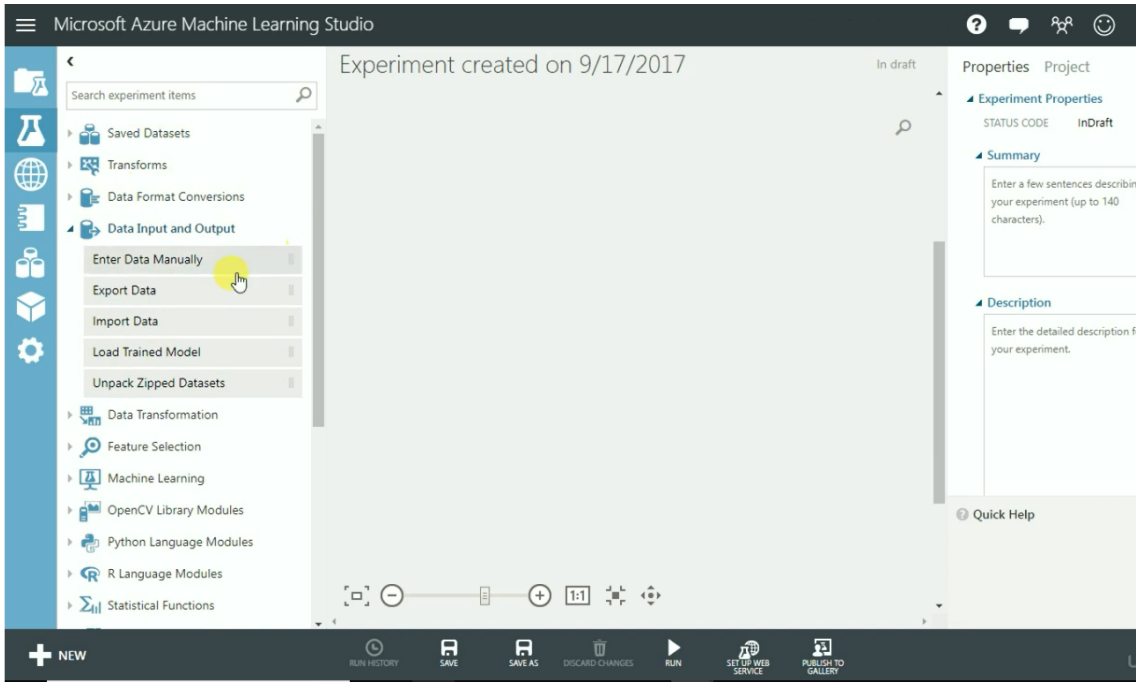


Delete the dataset and have some free space

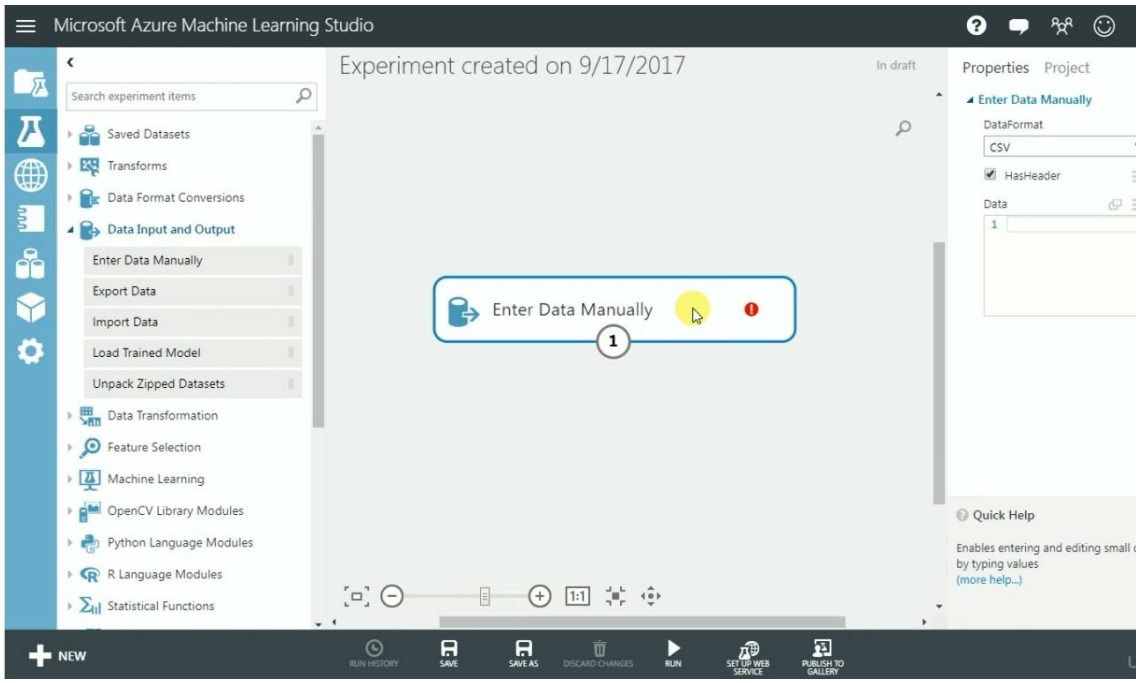


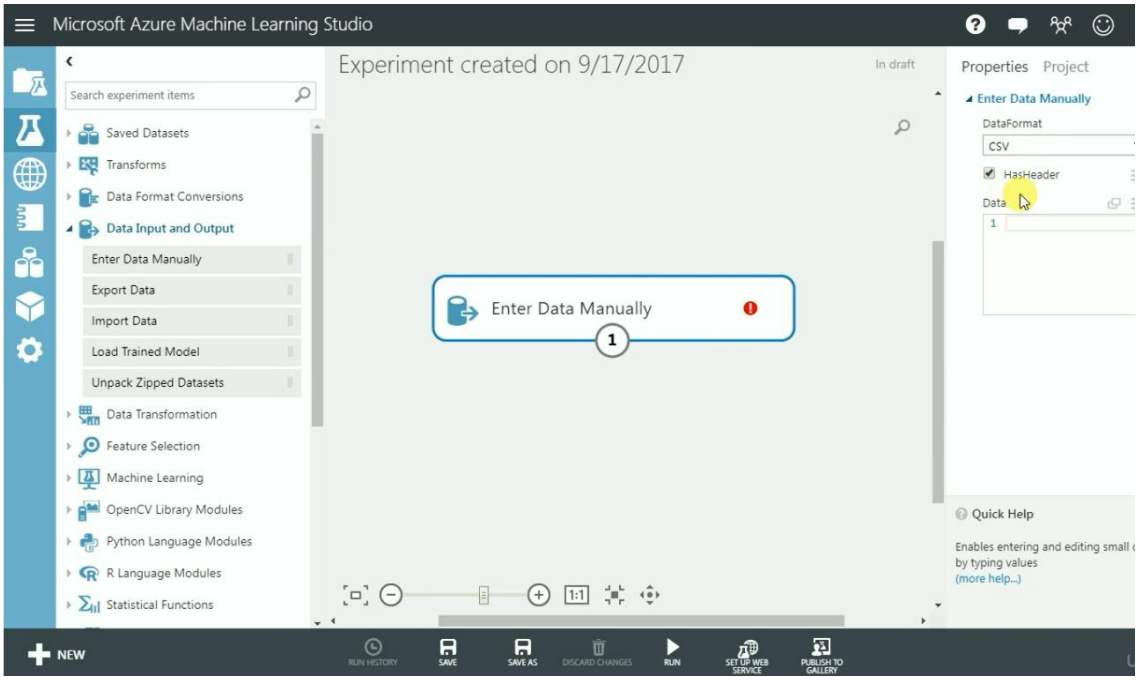
Entering Data Manually

Now we can see how to enter manual data

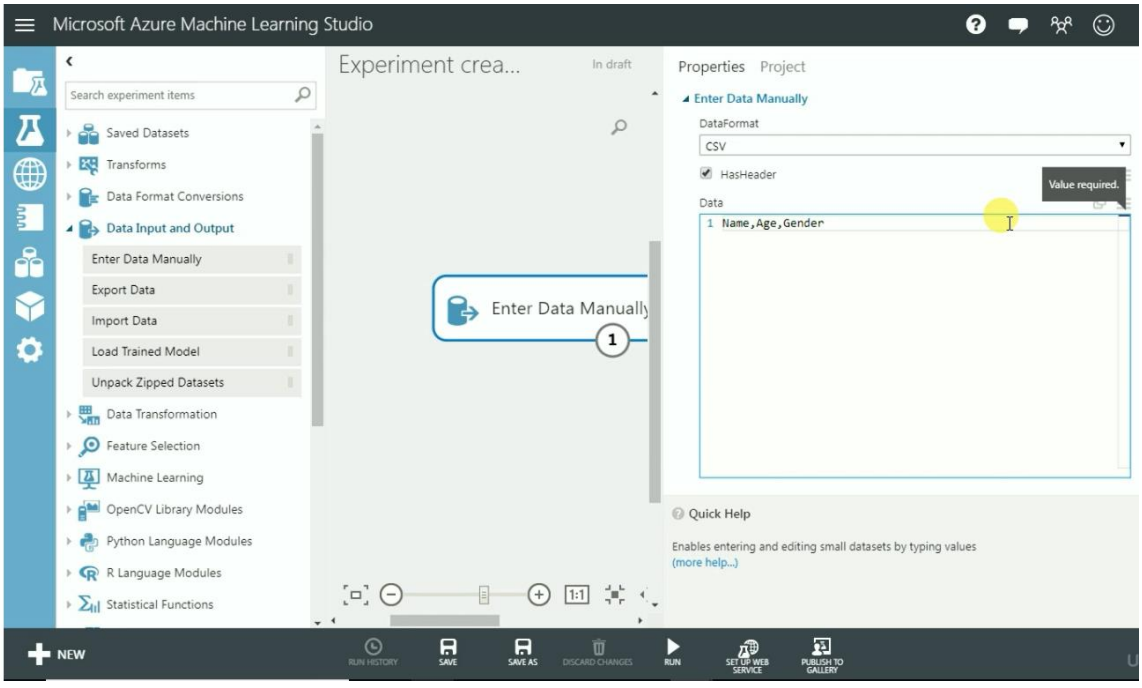


Drag and drop Enter data manually

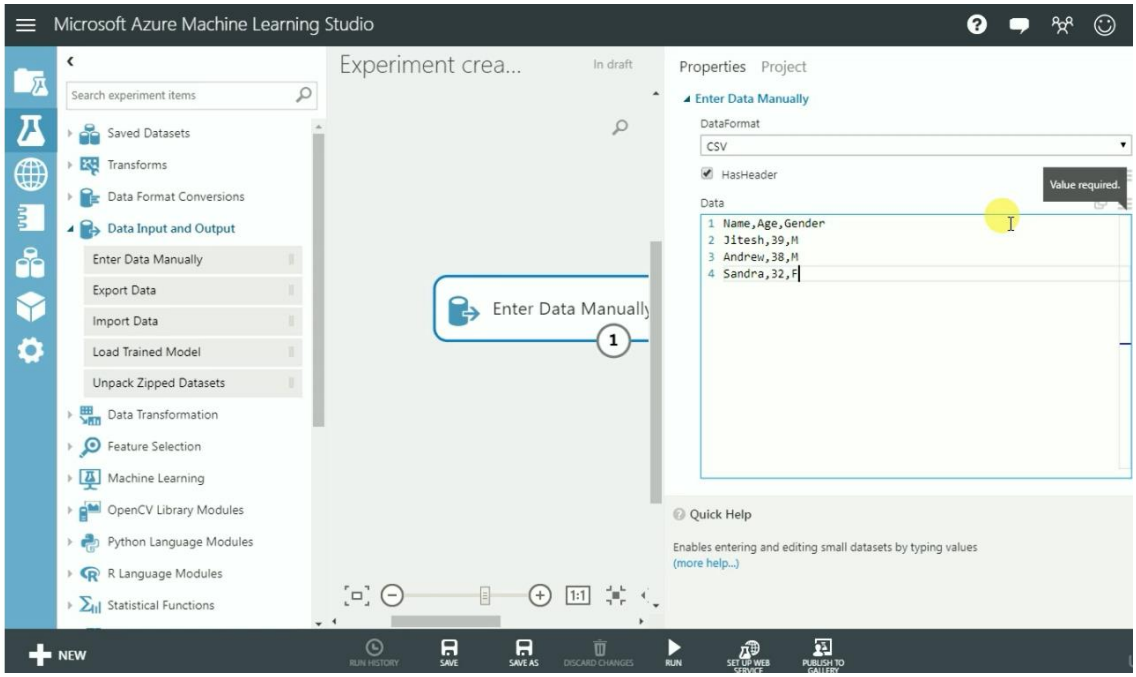




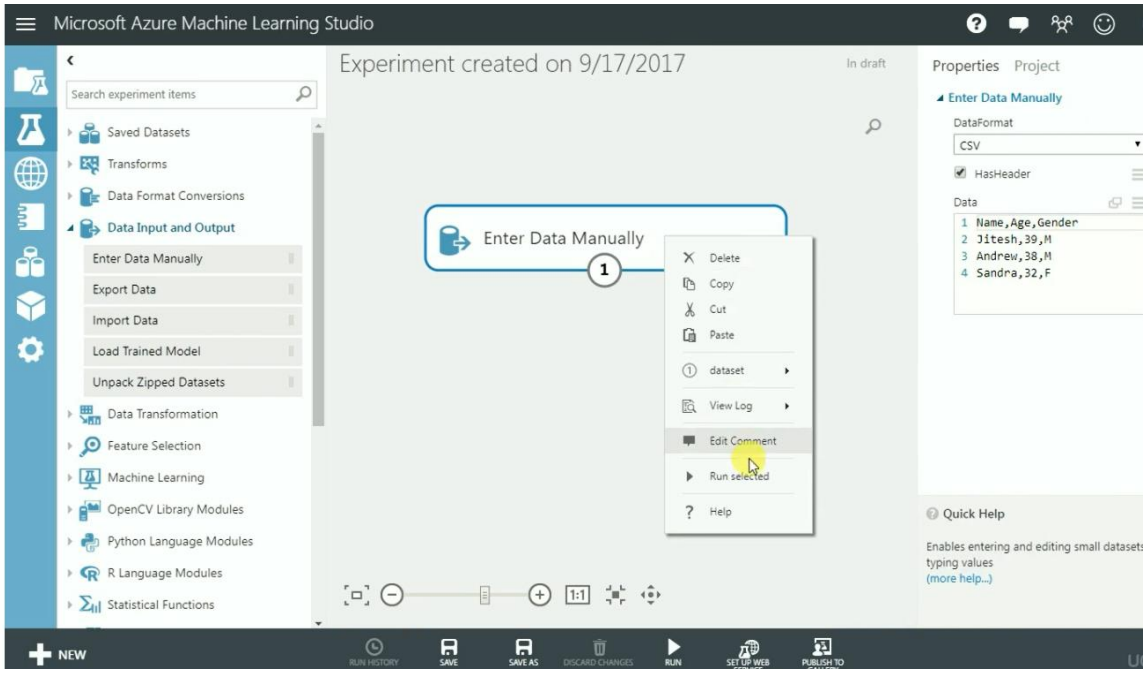
Type heading as shown below



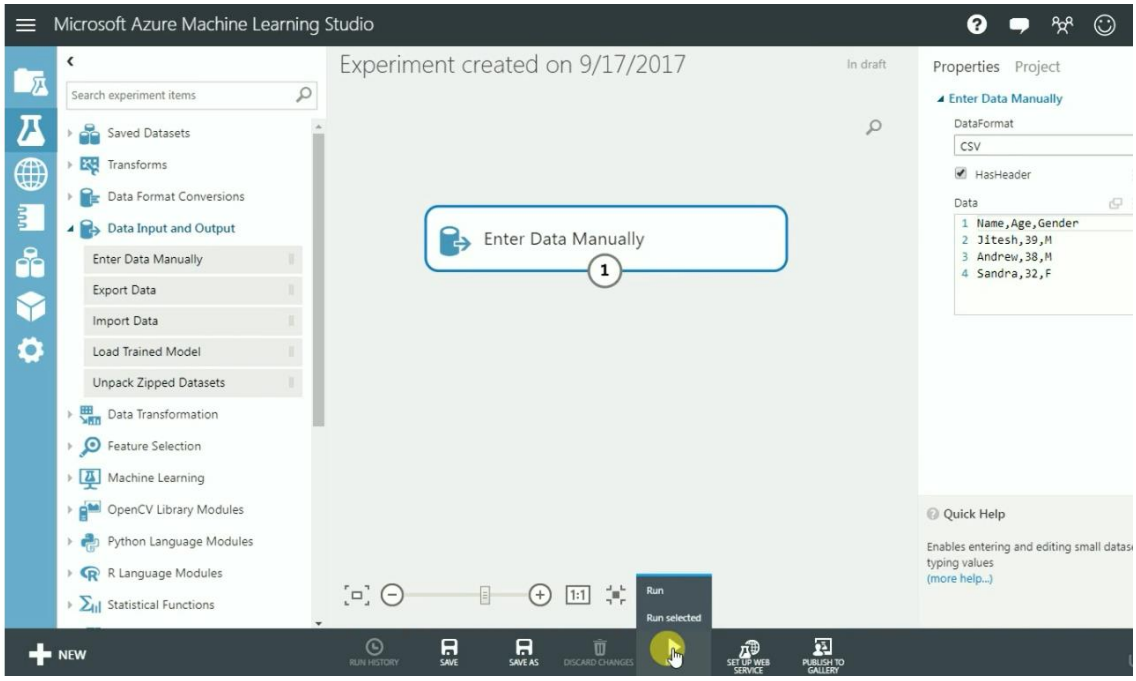
Type the data one by one



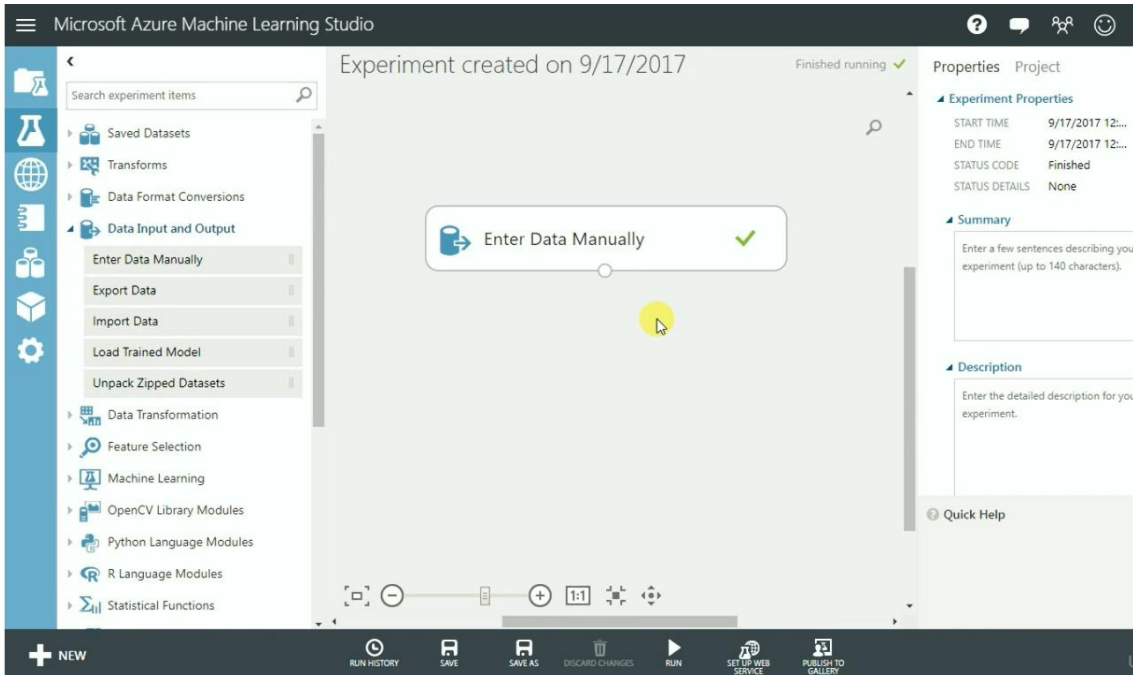
To run- right click and click run selected



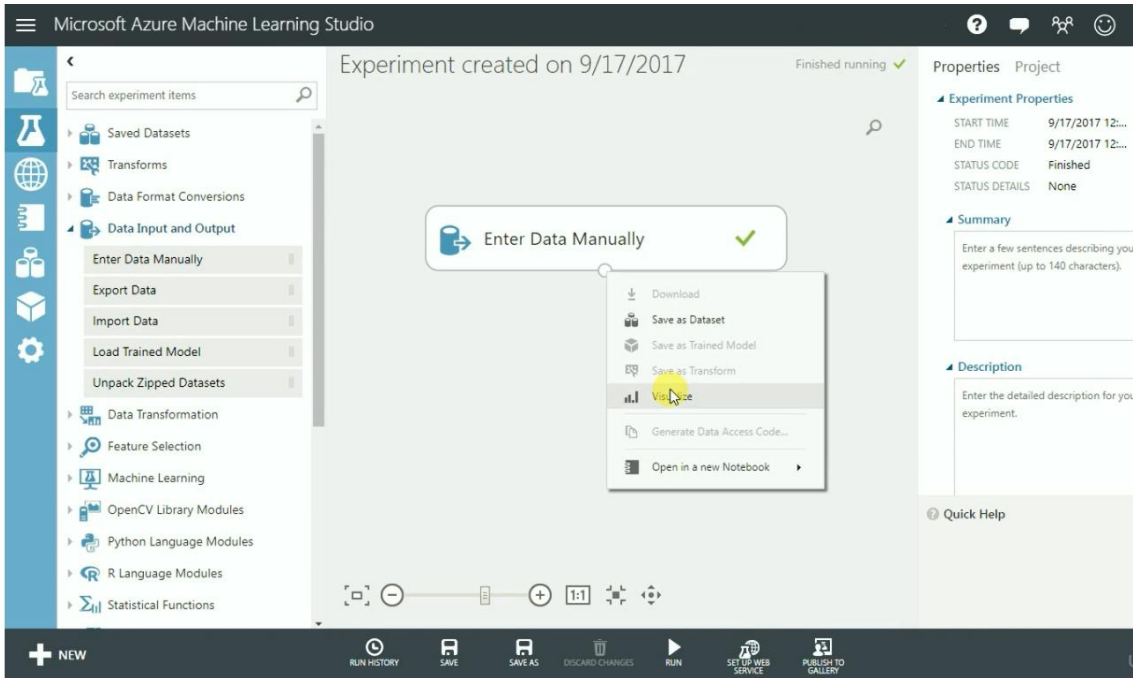
Or simply click run in the bottom



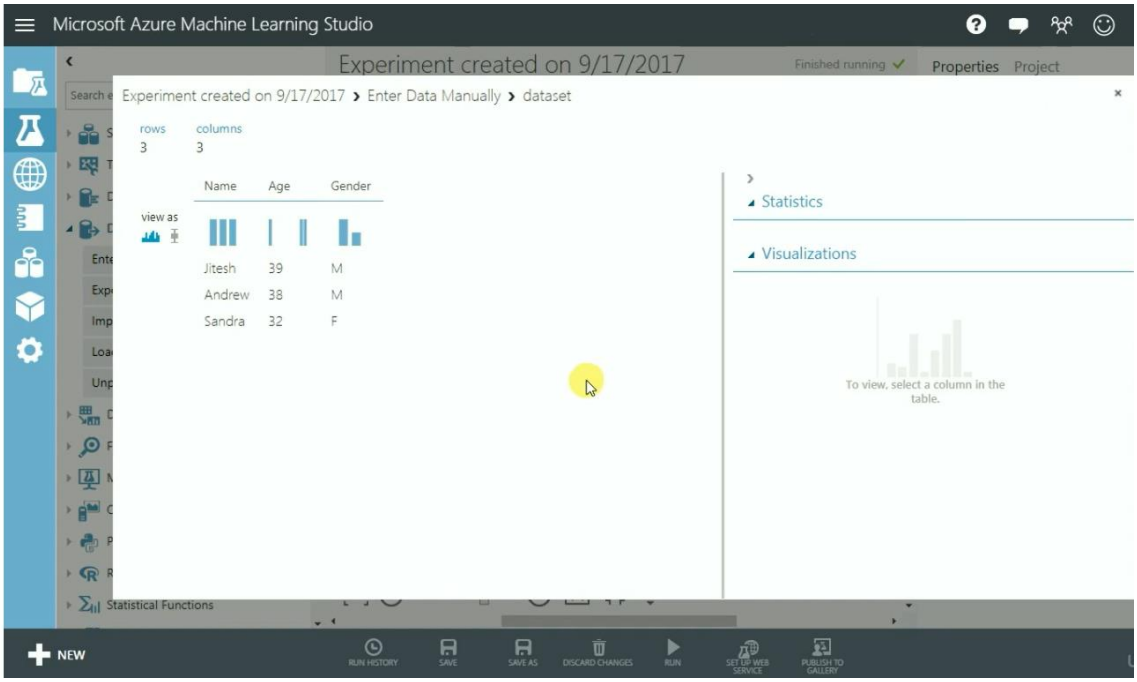
Finished running



Right click and visualize

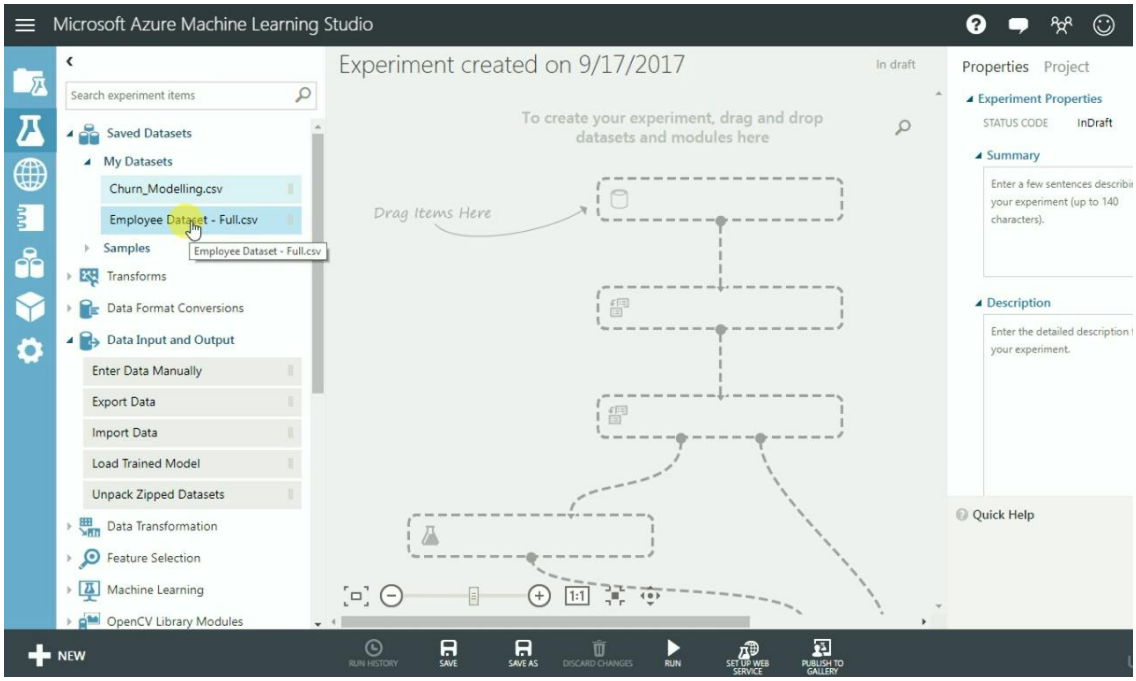


Created dataset with specified inputs successfully

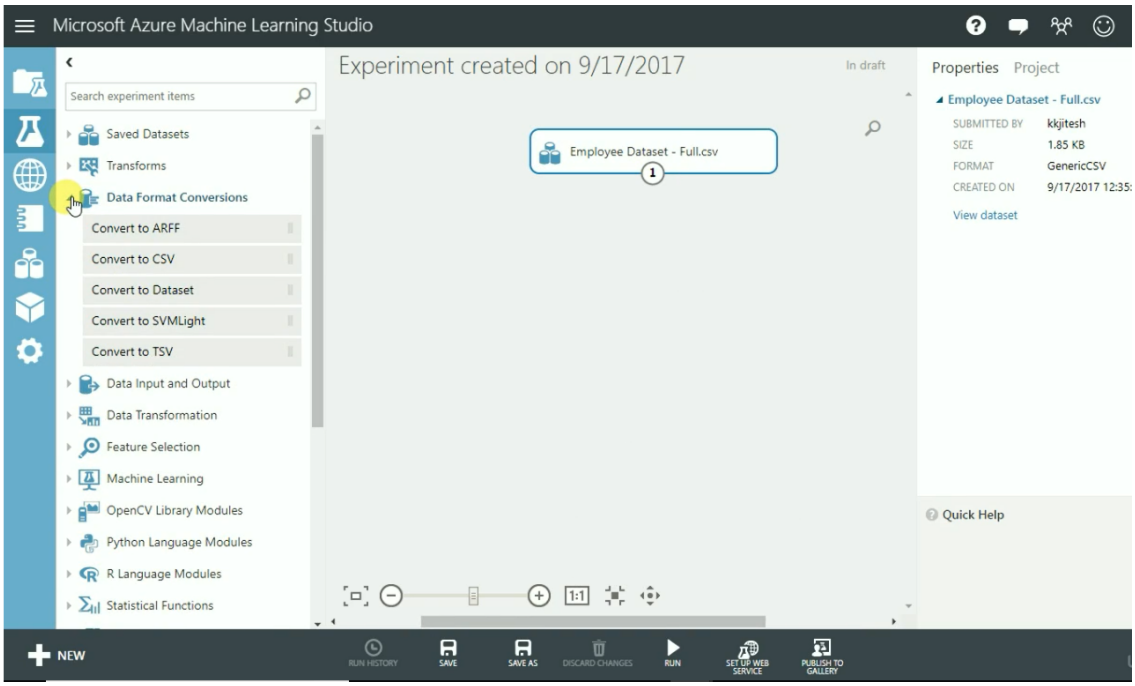


Converting File to TSV

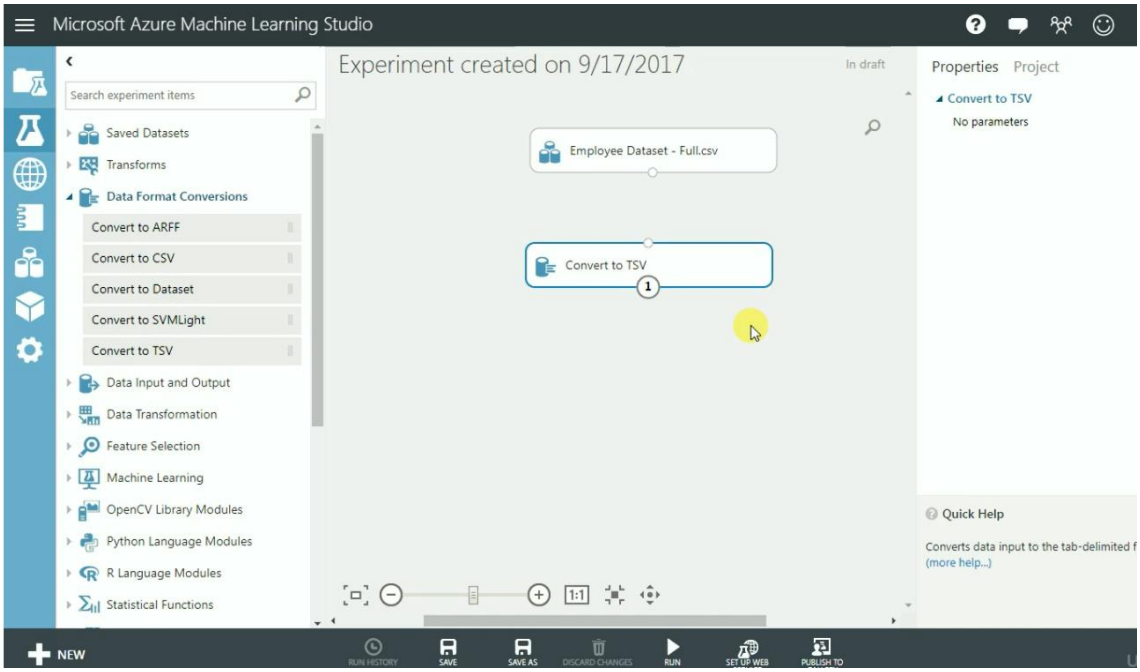
Pick dataset created earlier



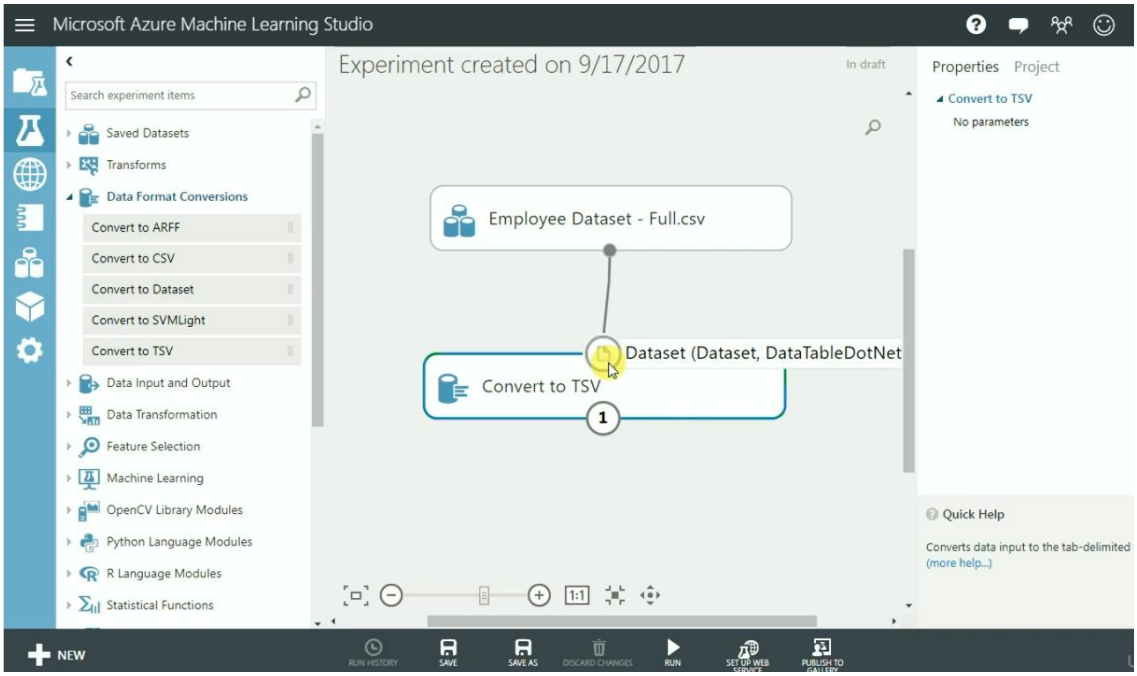
Drag and drop the dataset and click on data format conversion



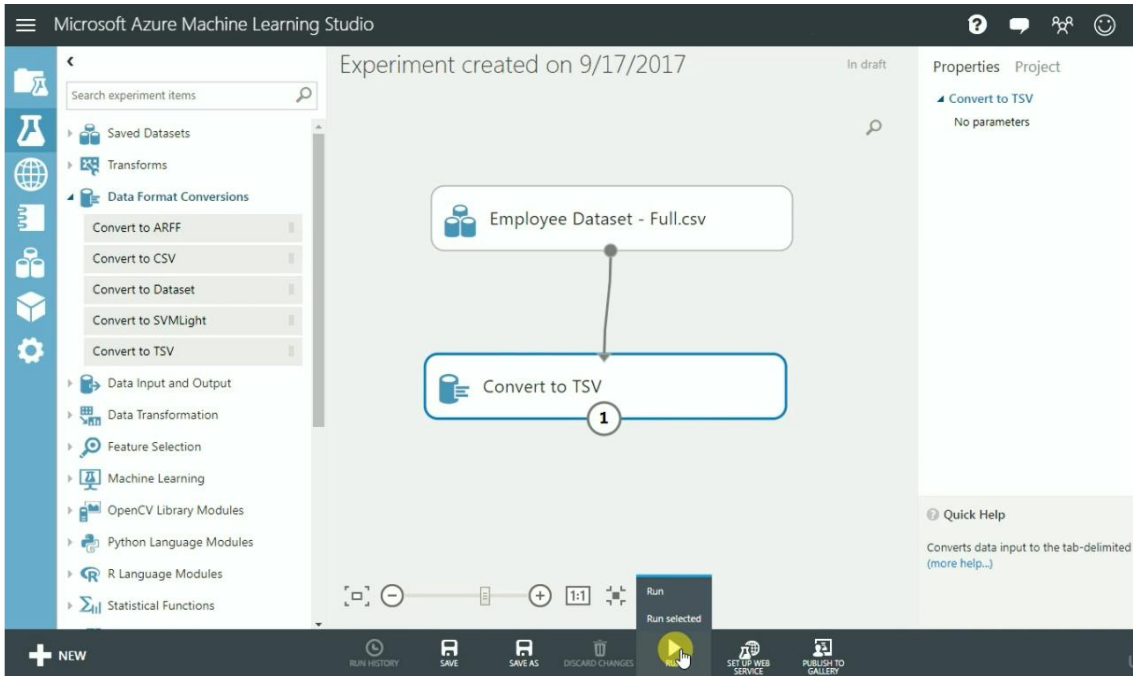
Drag and drop convert to TSV module



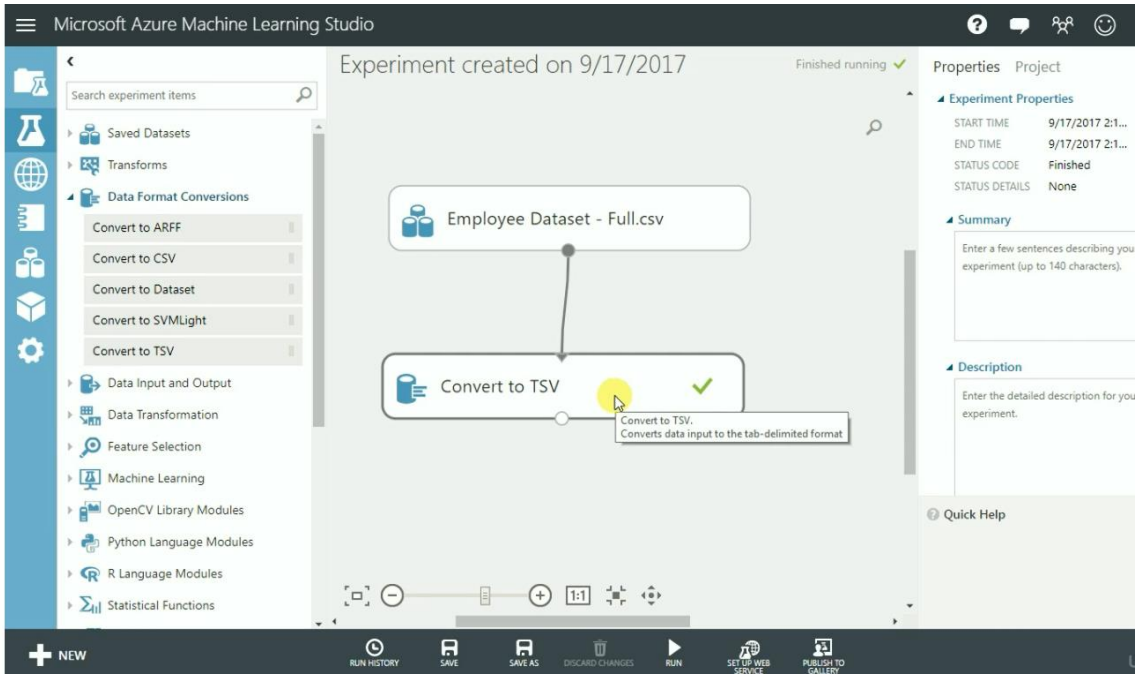
Simply attach Output node from CSV to TSV as shown



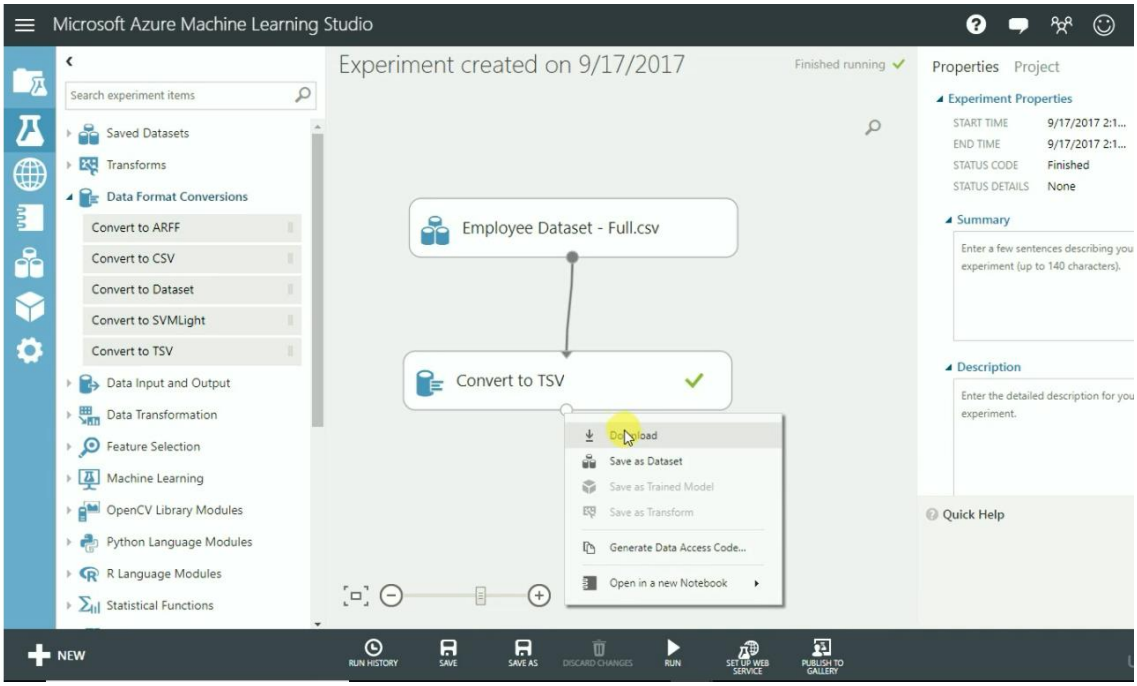
Run the module for execution



Now Azure has converted from csv to tsv file.

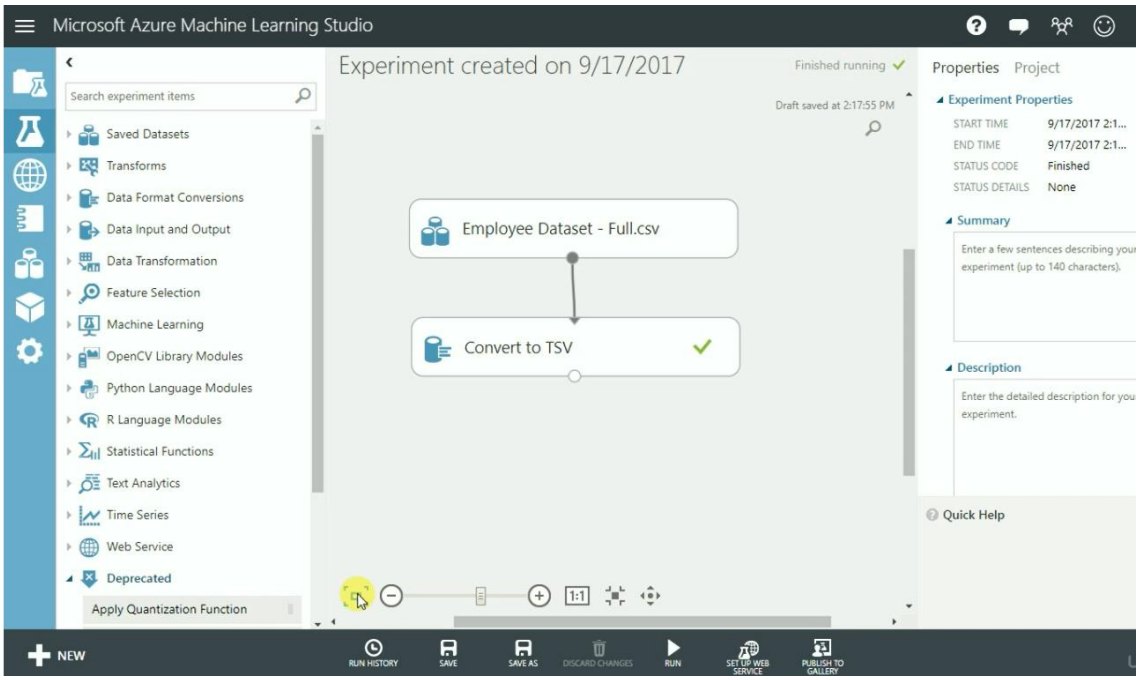


Can download the same by right click and download to check the output

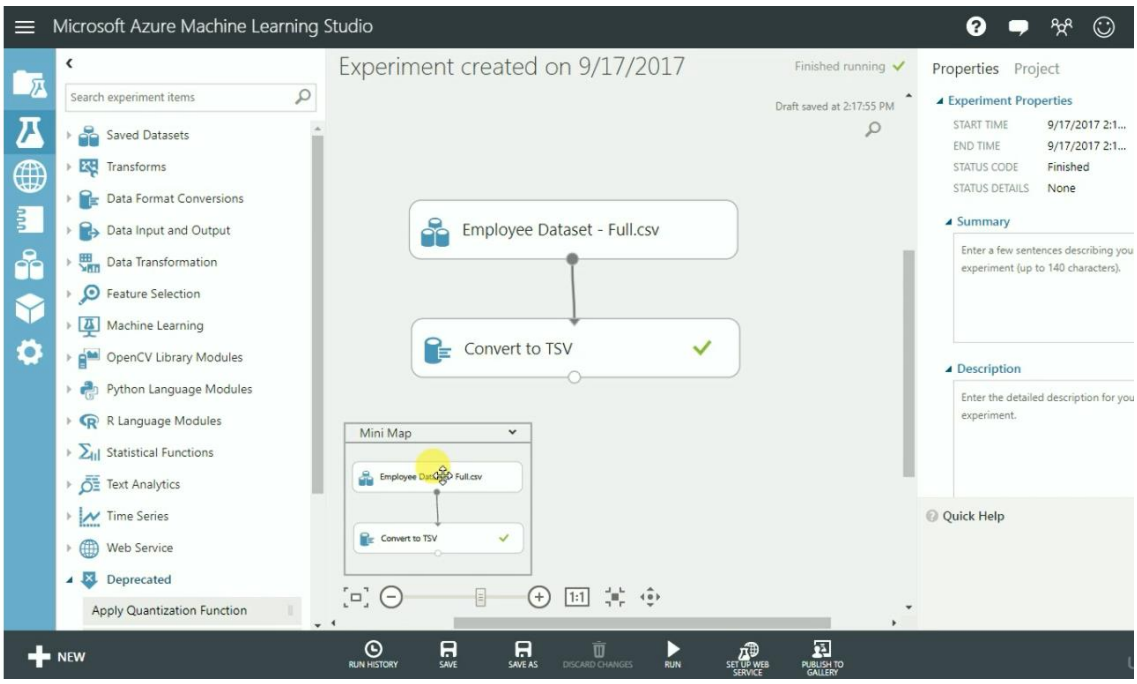


Understanding on Mini View

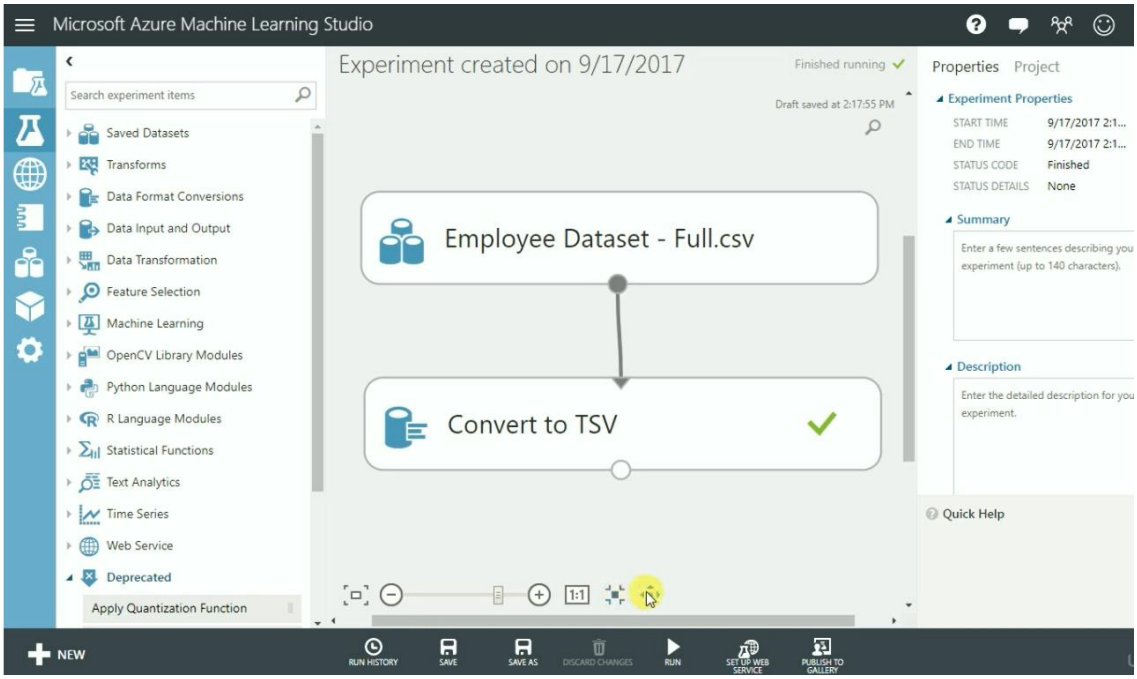
Click on mini map to preview the module



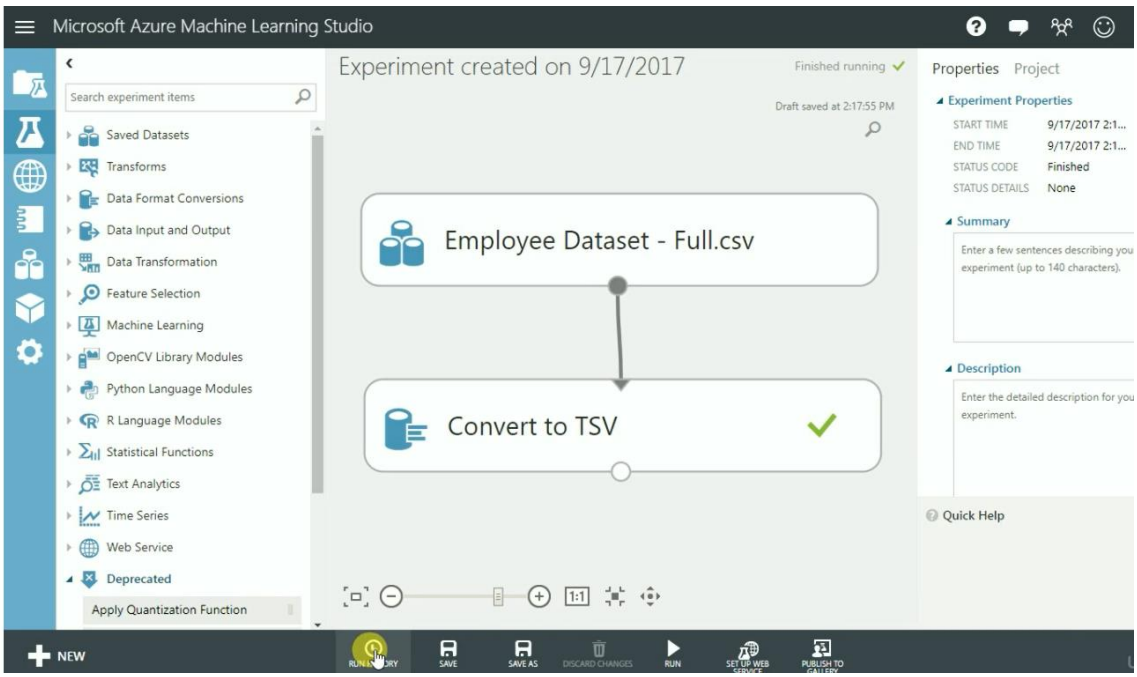
Can preview module as below using mini map

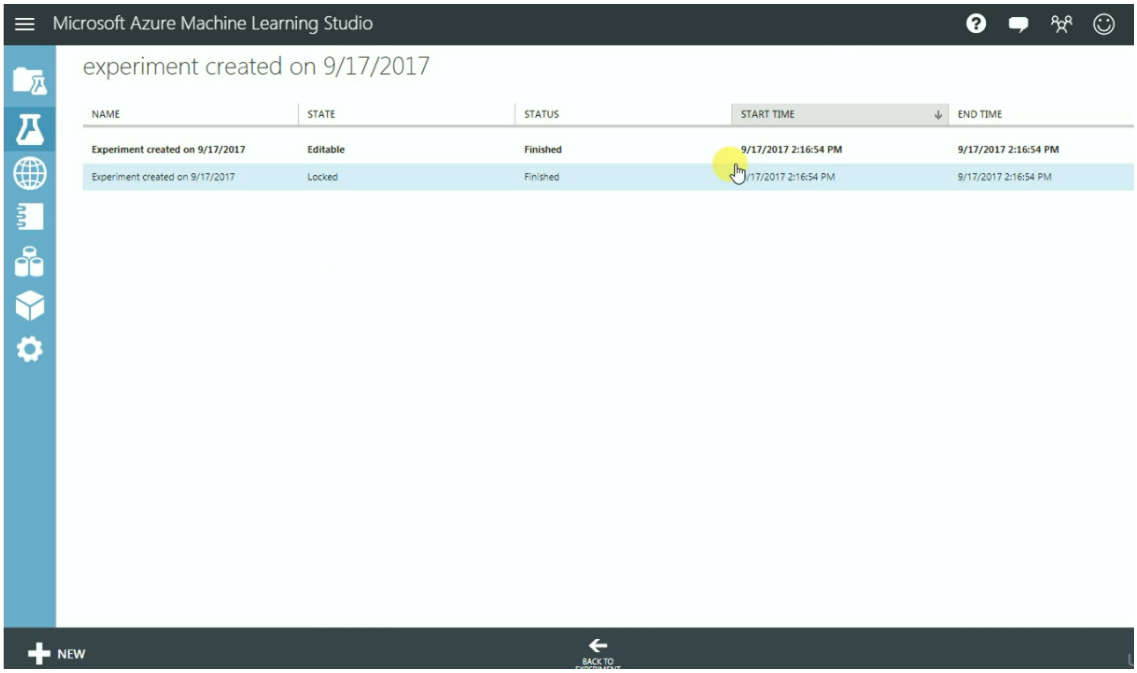


Can Perform actions like Zoom in, zoom out and make fit to screen click below

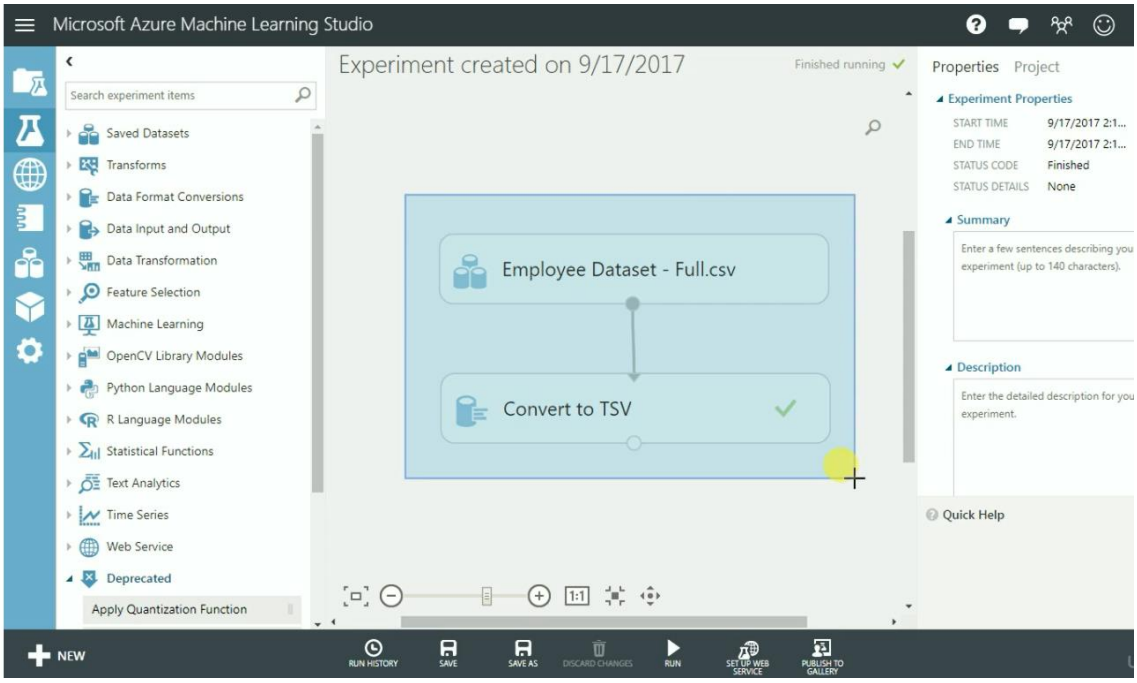


By clicking run history, can view experiment history

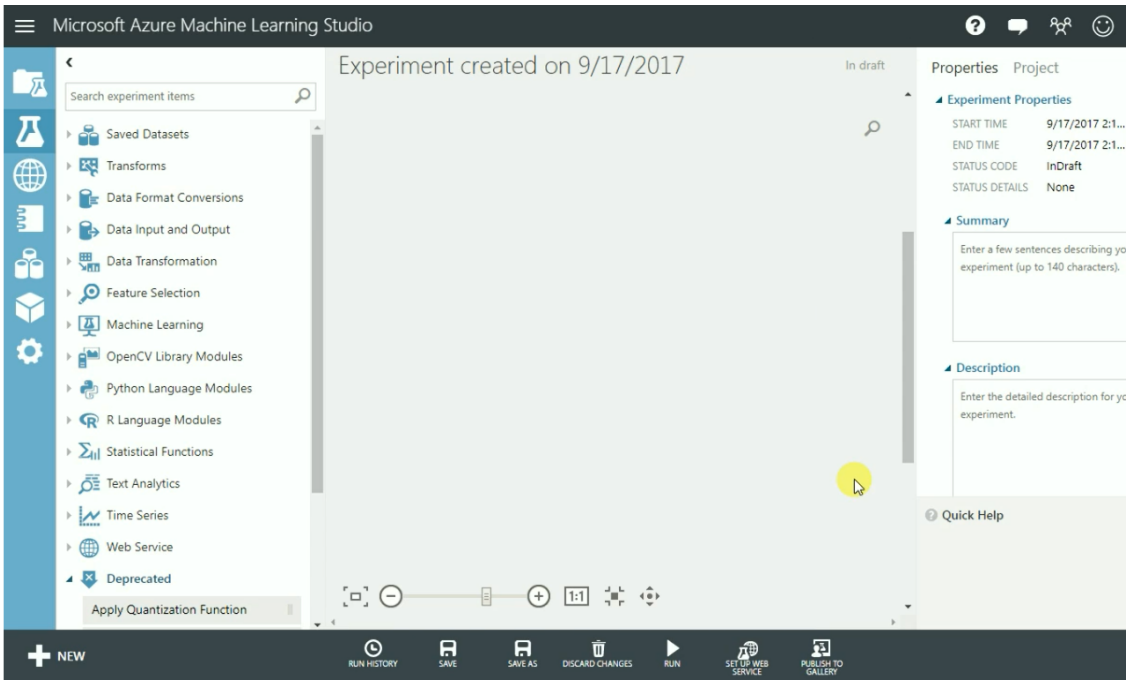




To delete the modules simply select the modules and press delete for free space

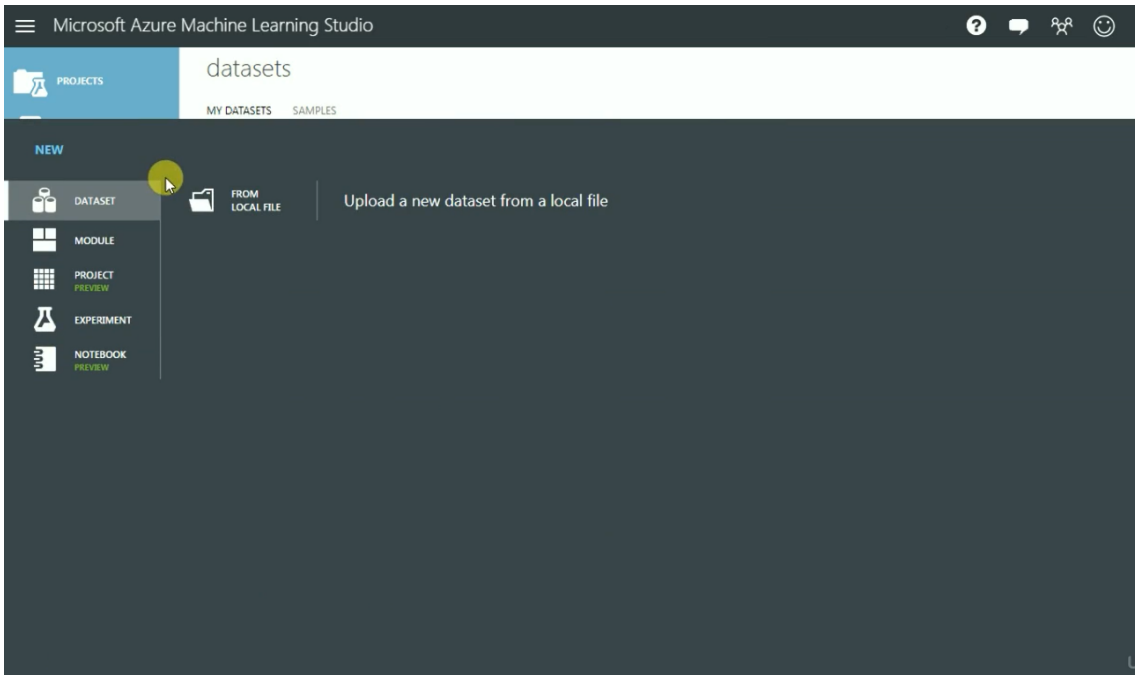


After deletion

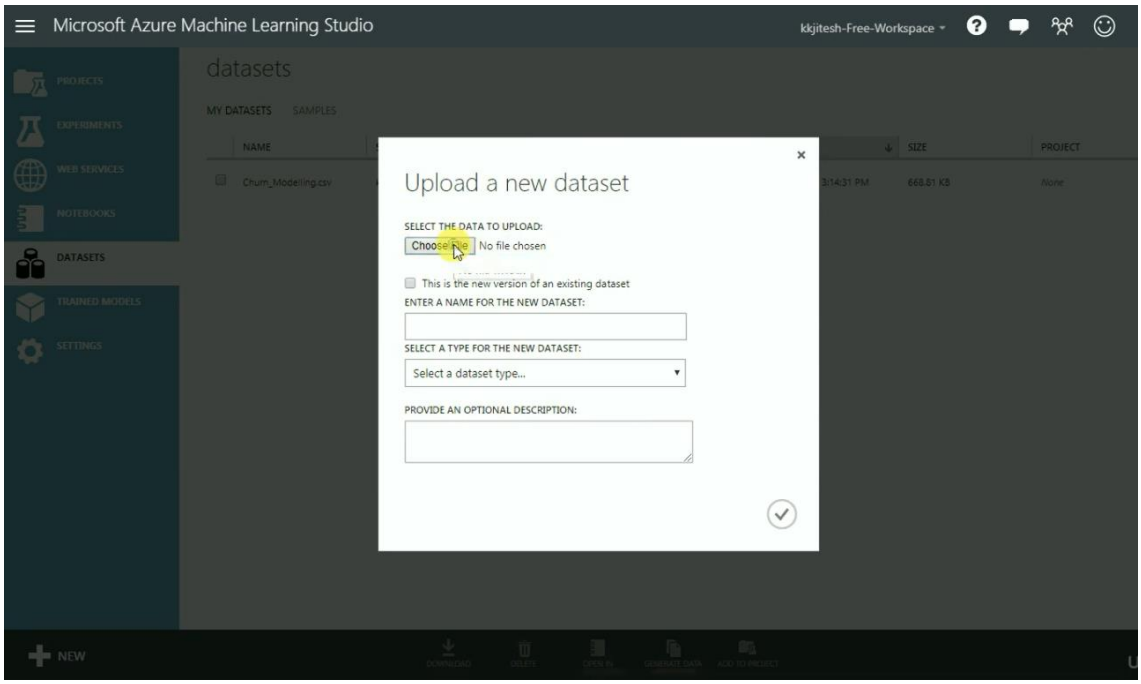


Unpacking the Zipped Dataset

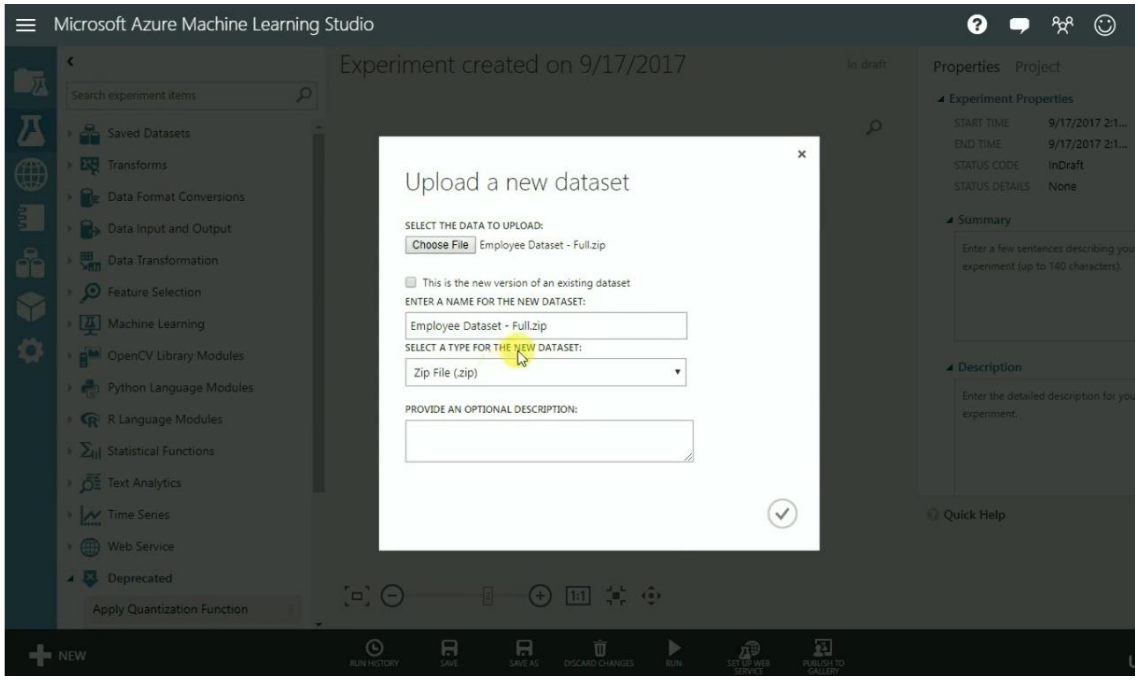
Go to dataset and click From Local file



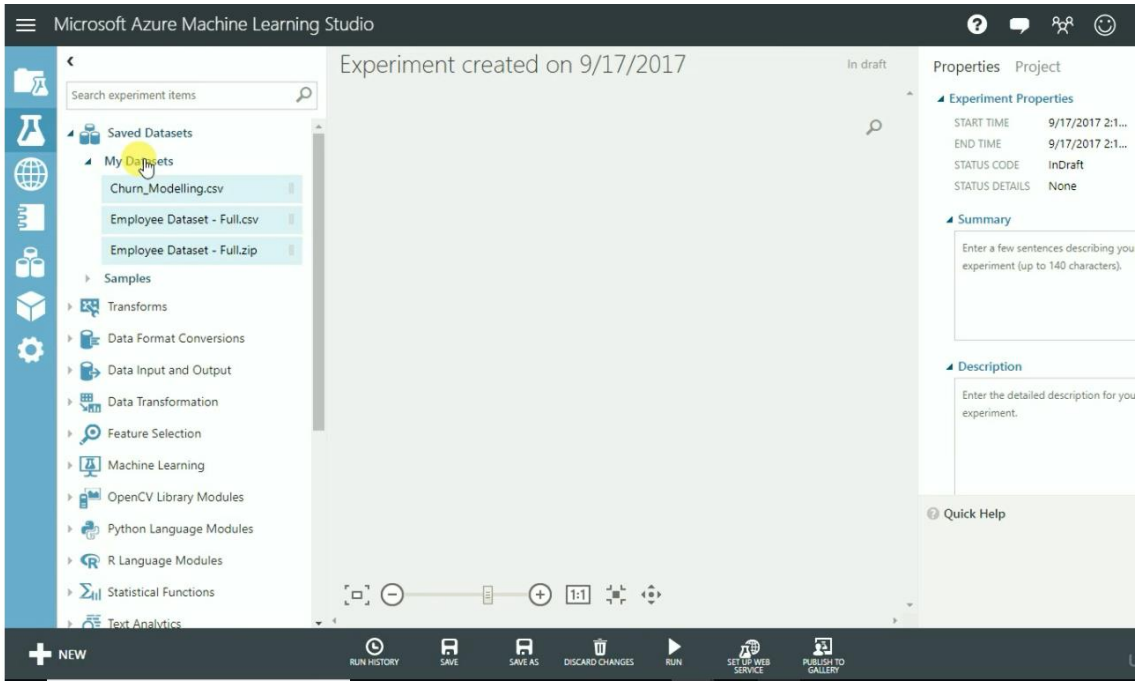
Choose zipped file from local drive



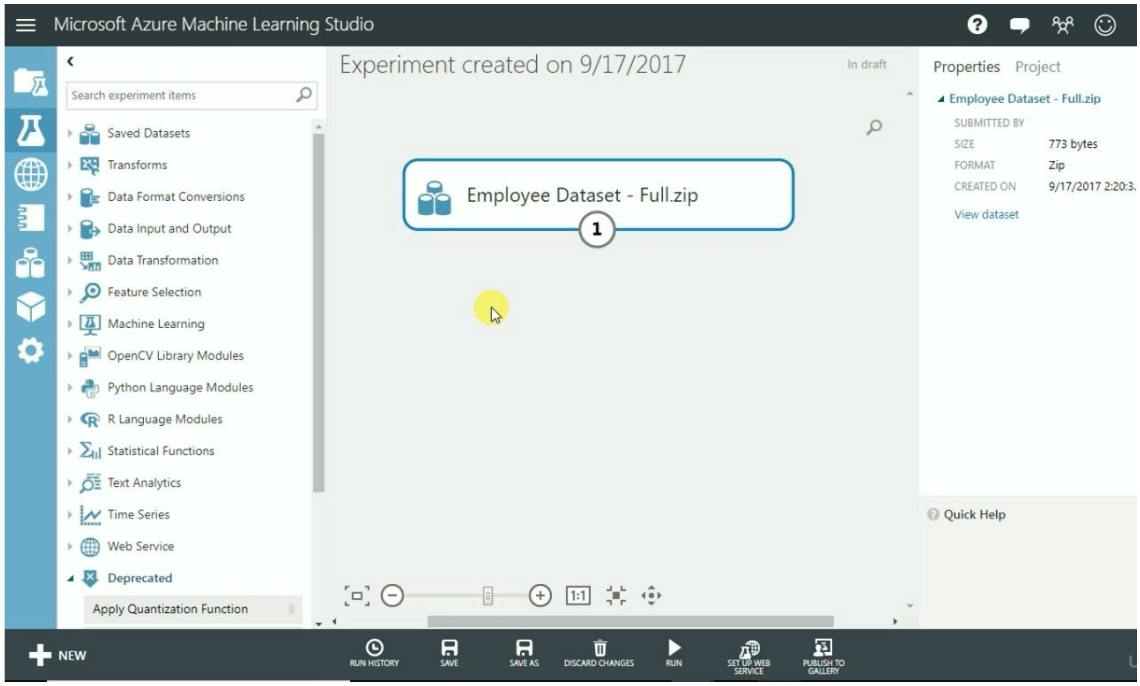
Upload the zip file



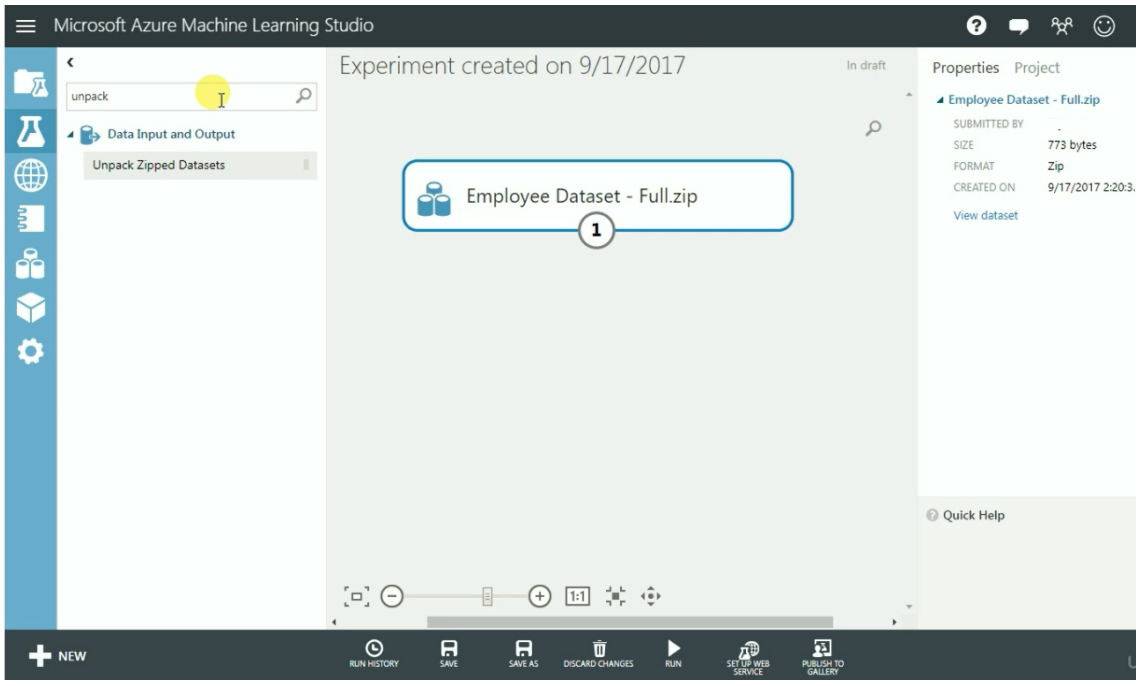
Click on saved datasets→ My datasets→ Zipped file



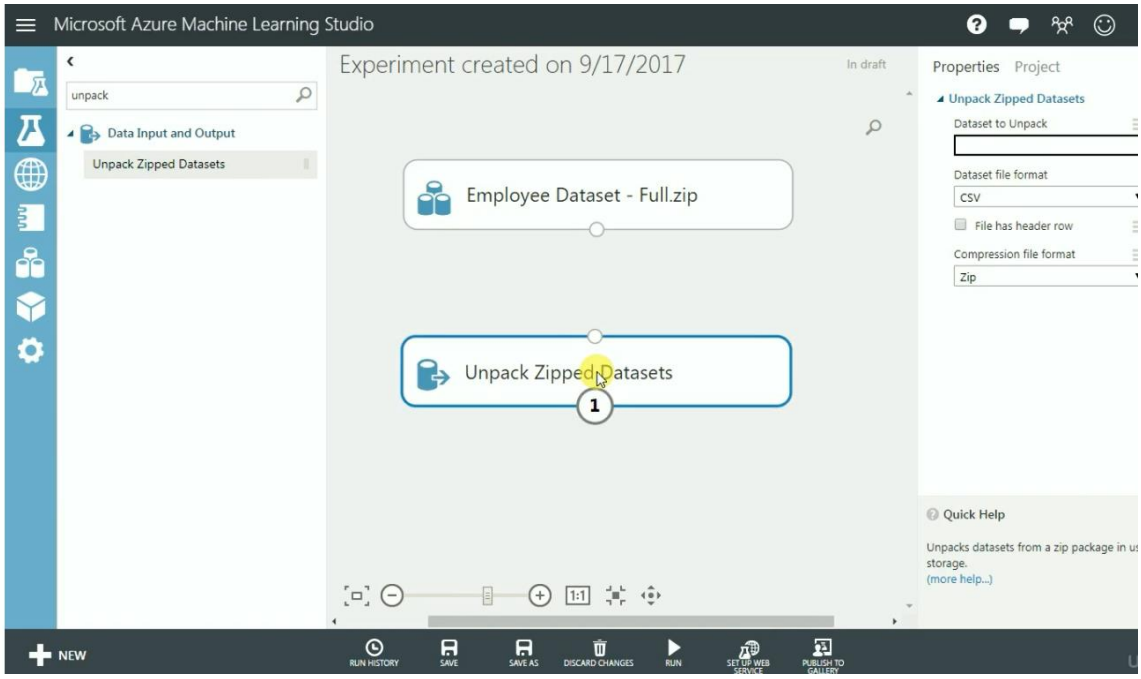
Drag and drop the dataset



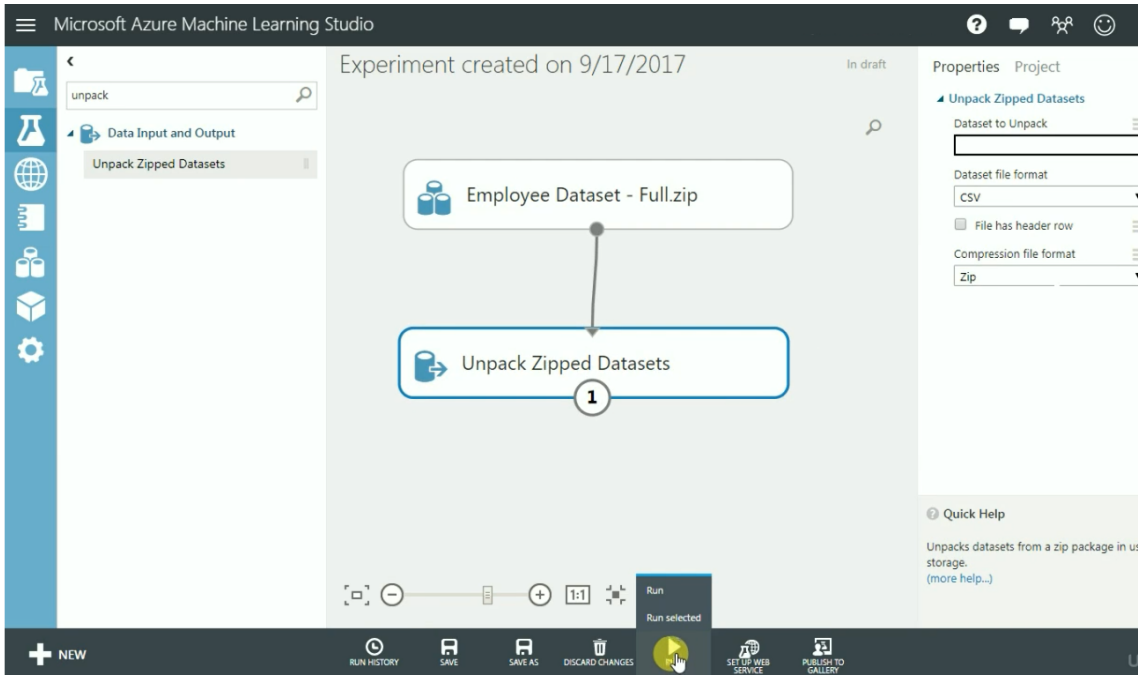
Search for unpack Zipped datasets



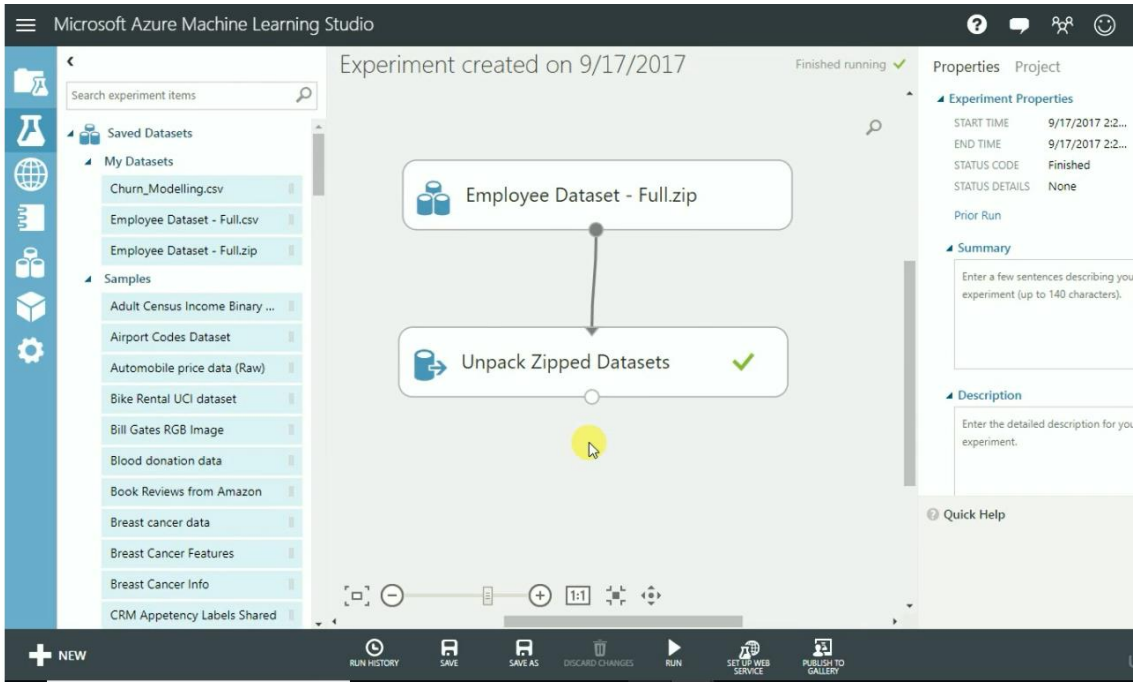
Drag and drop Unzip dataset



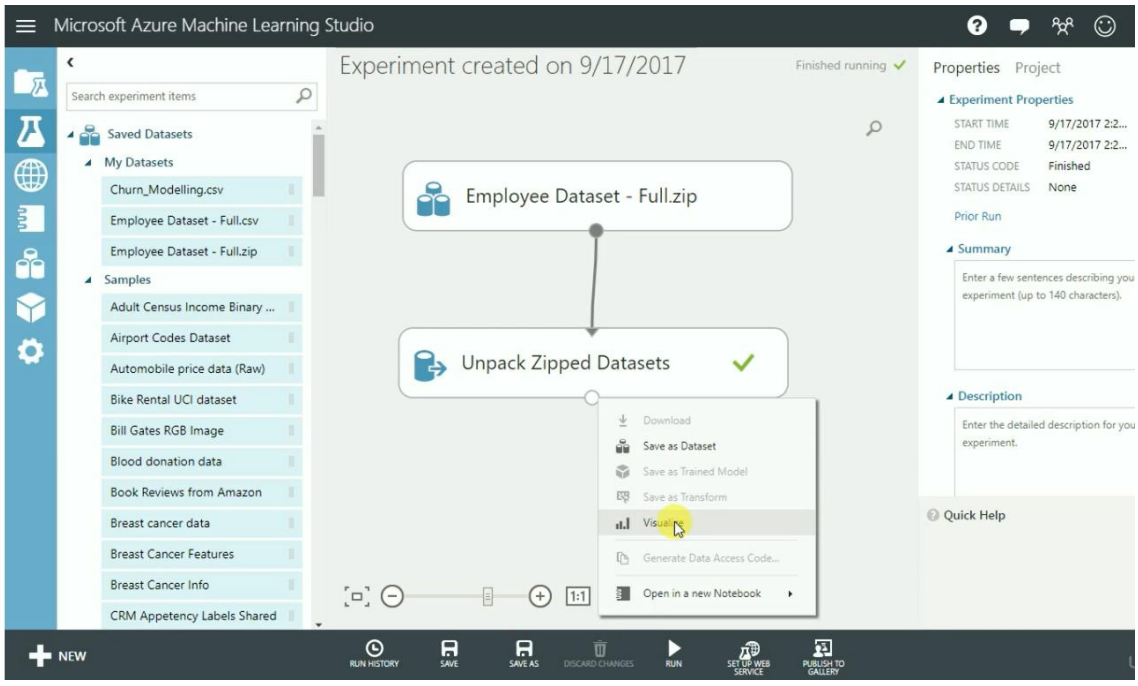
Connect the output and input nodes and run the module



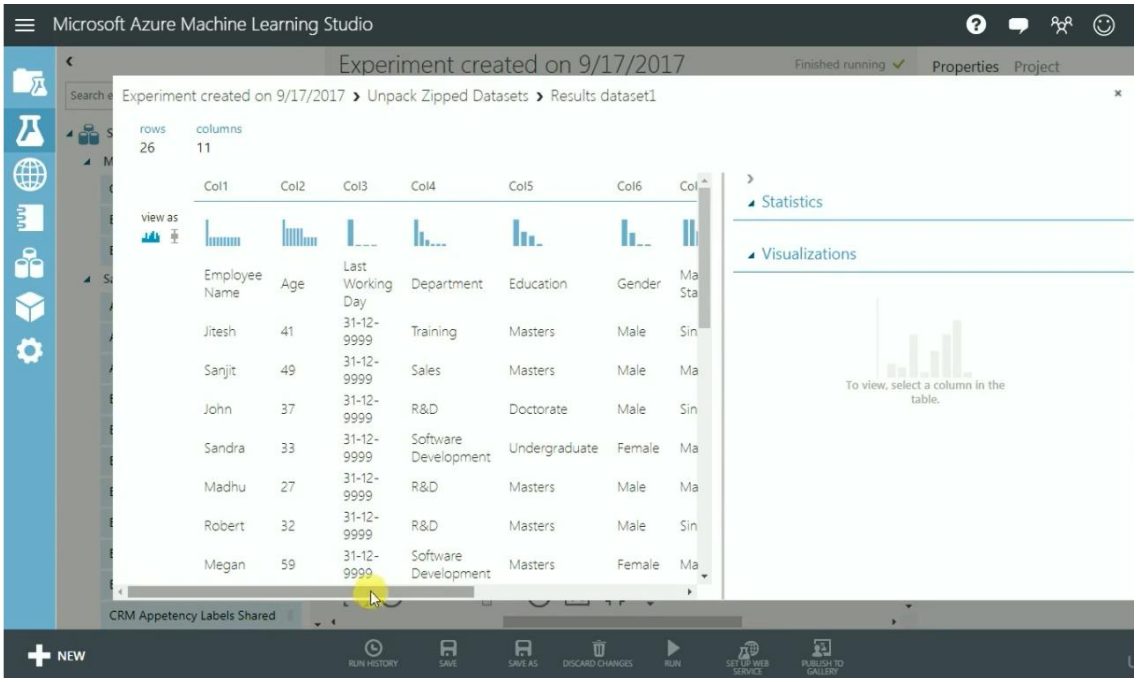
Post successful execution



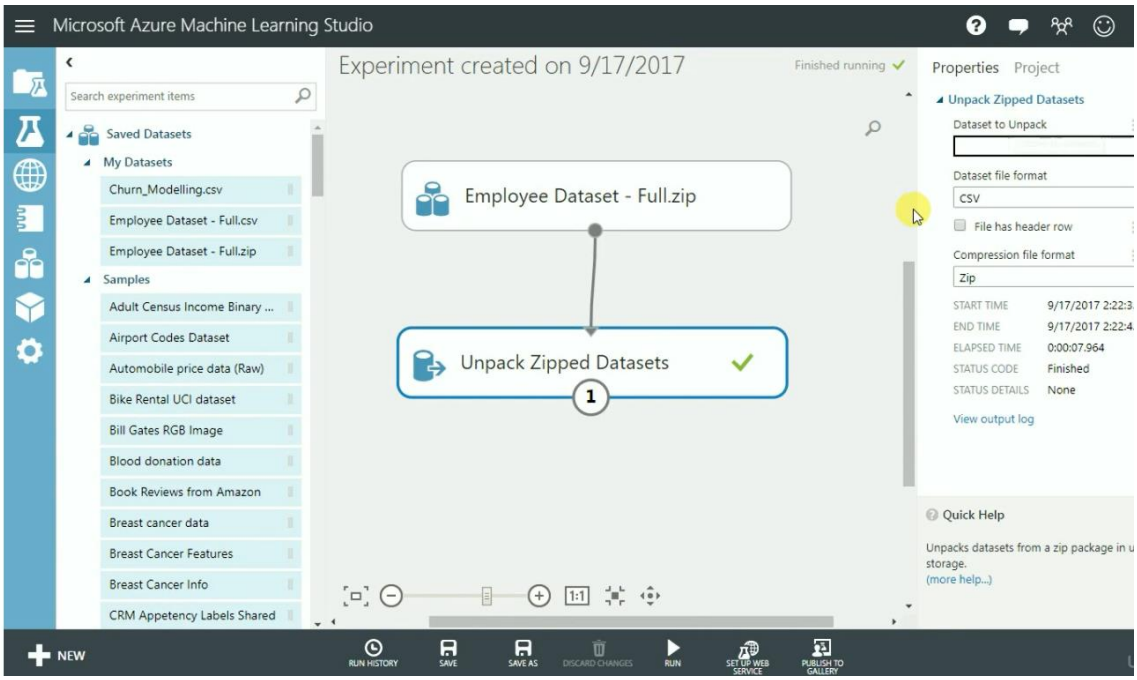
Right click on output node and visualize



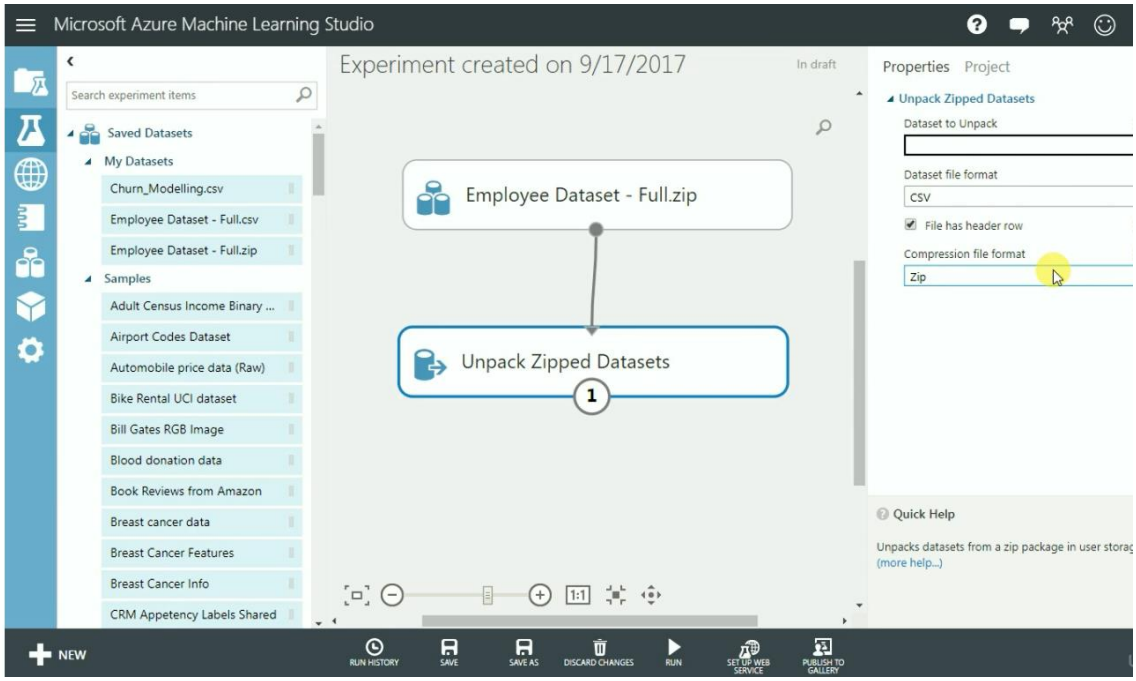
Visualize and check the output. However, result obtained header shows col1, col2 etc. to sort out this refer next slide.



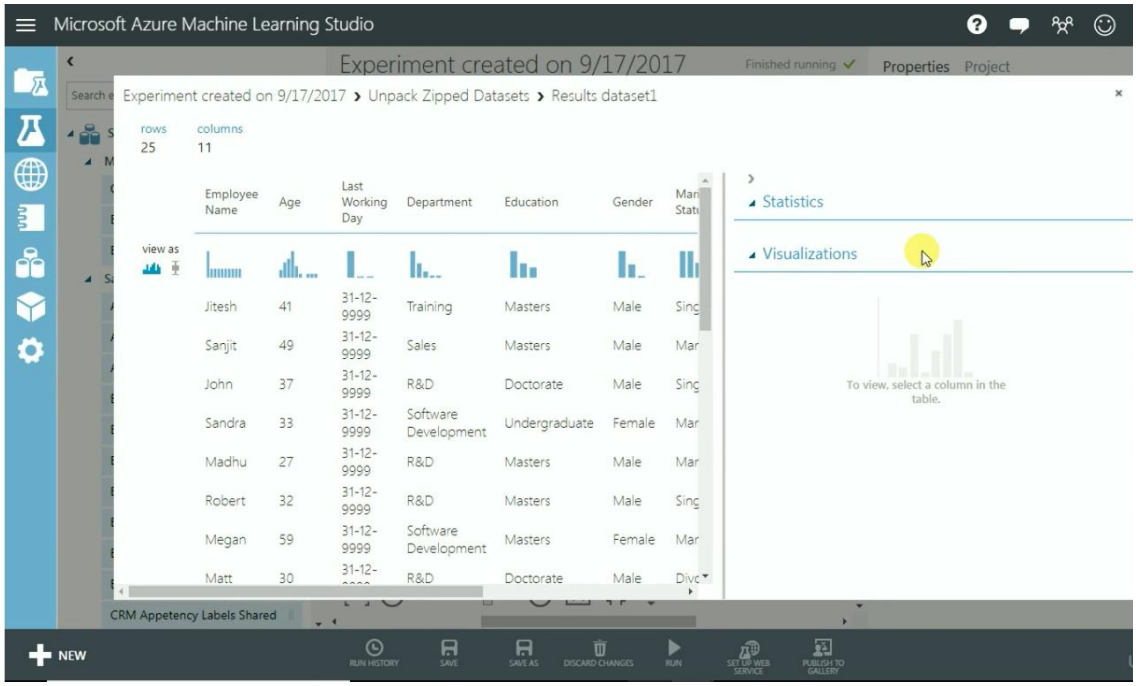
Click on file has a header row as shown in right side and run the module



Mark the file has header row and run the module again

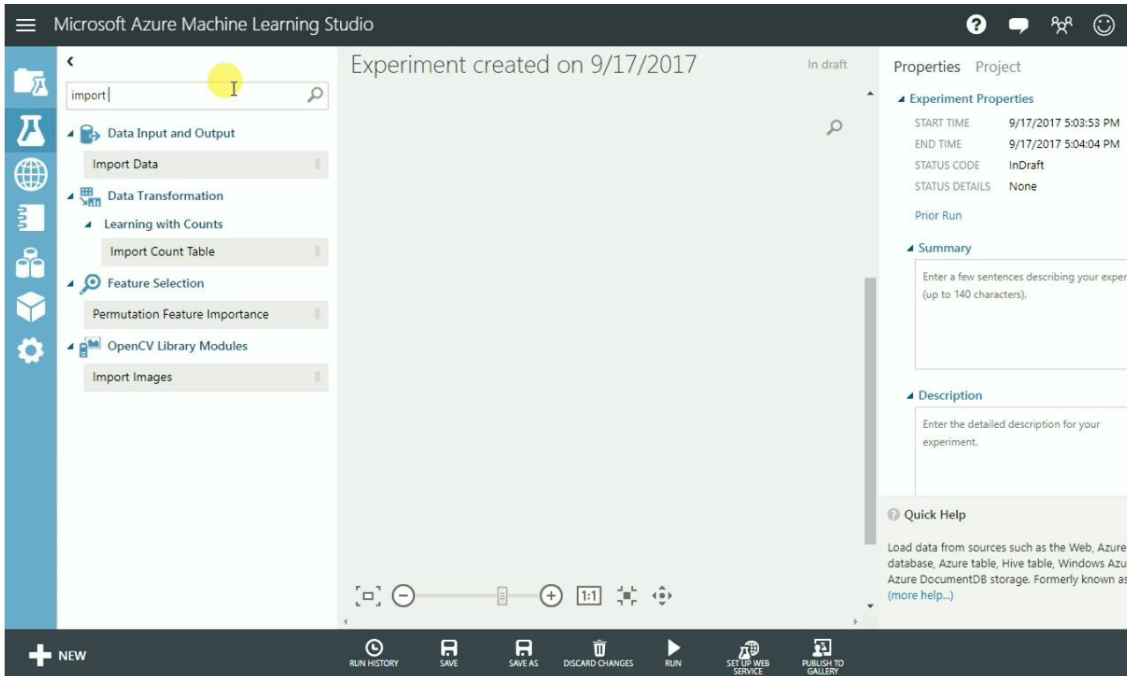


Post execution find the output obtained is correct

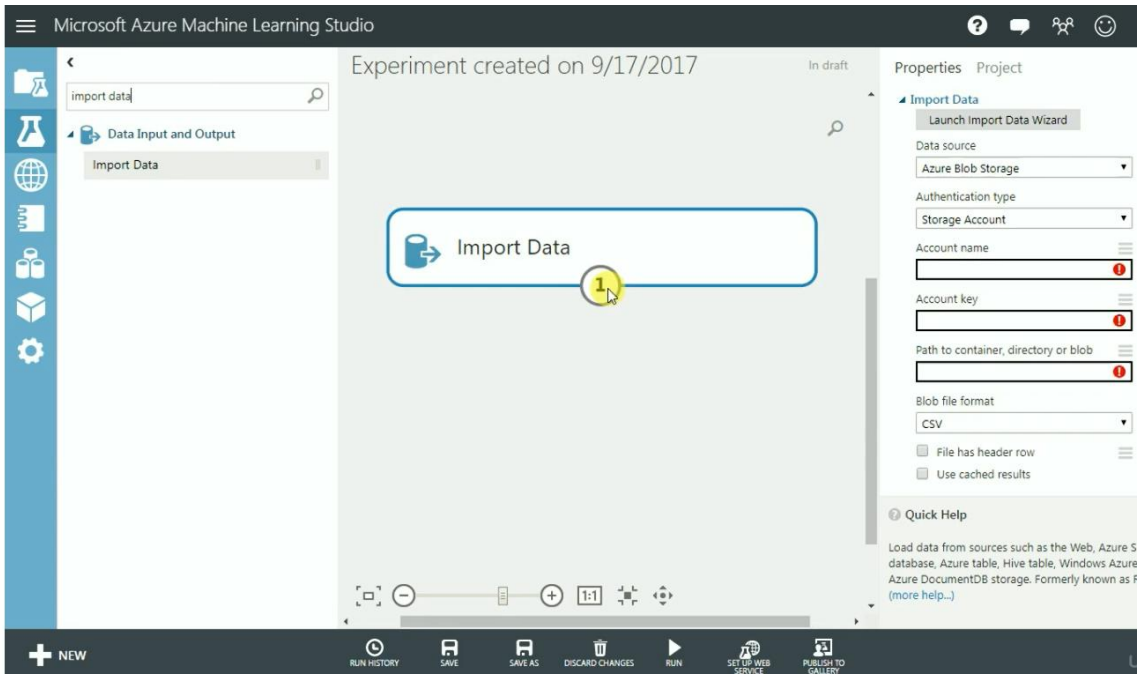


Importing Data from Multiple Sources

Search for the import data

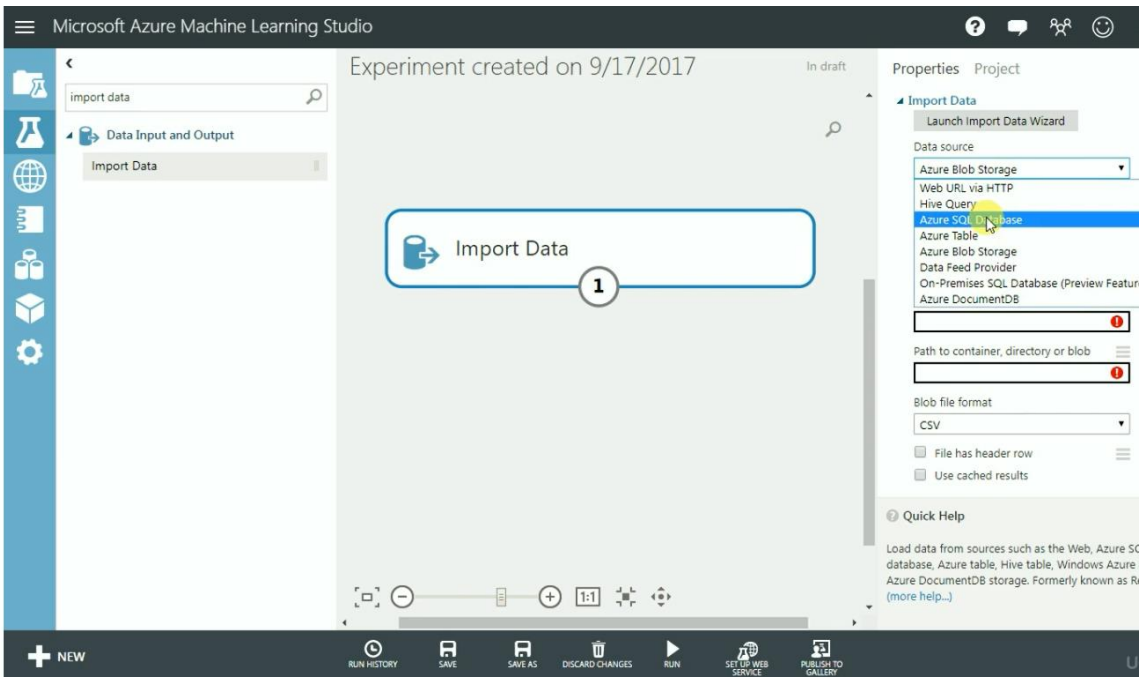


Drag and drop to the canvas



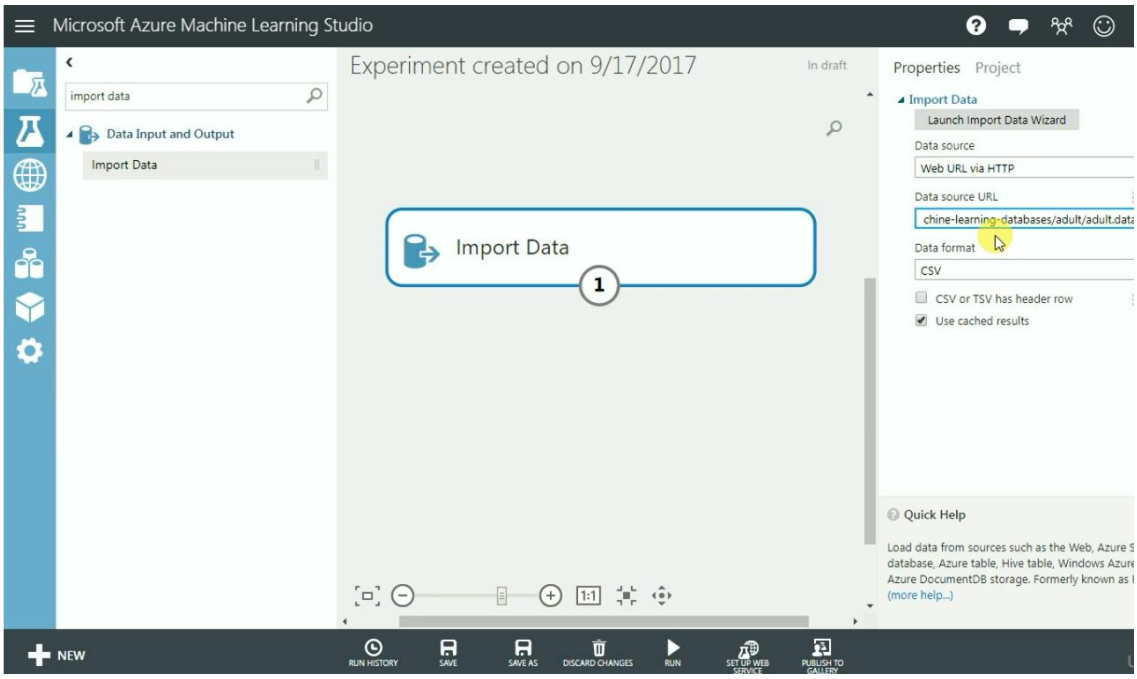
Can import data from various sources as shown below

Select Web URL via HTTP from the dropdown

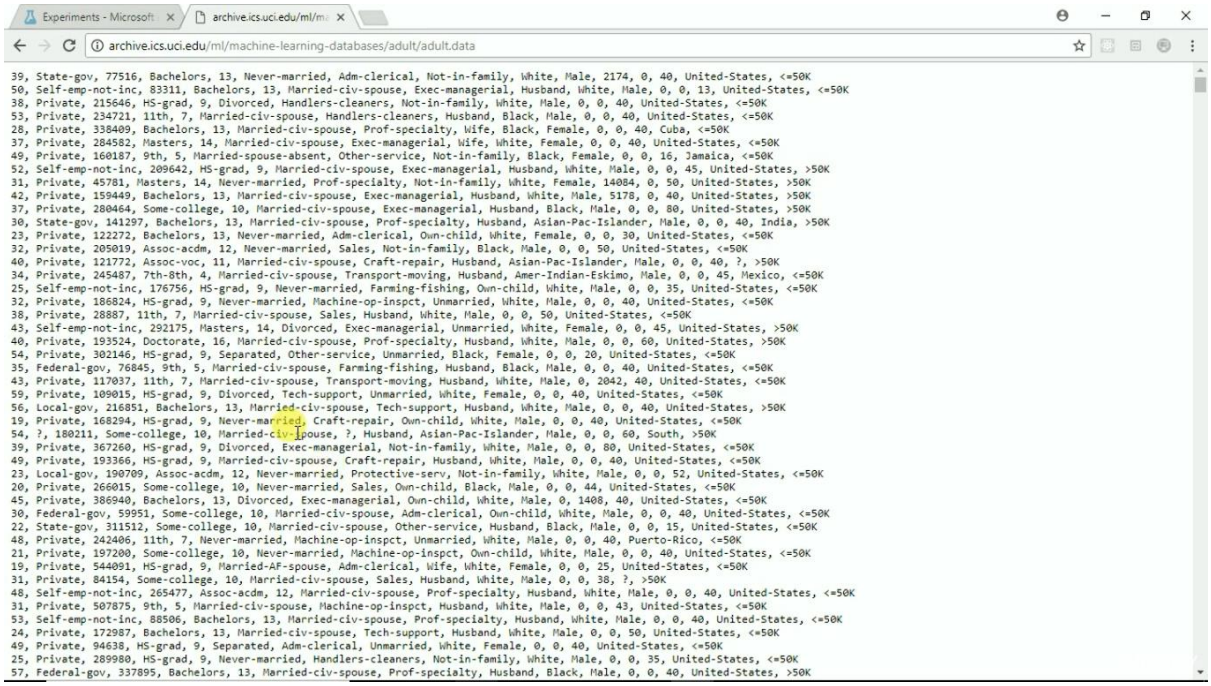


Input the URL from where data to be imported in next column

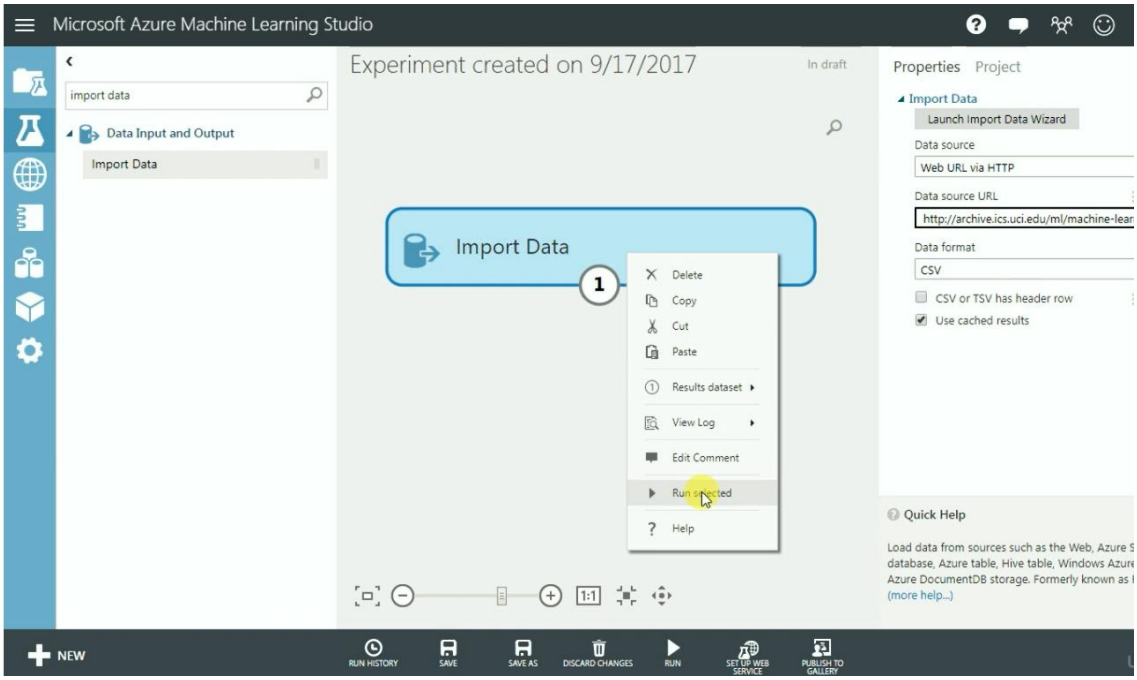
URL : **https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data**



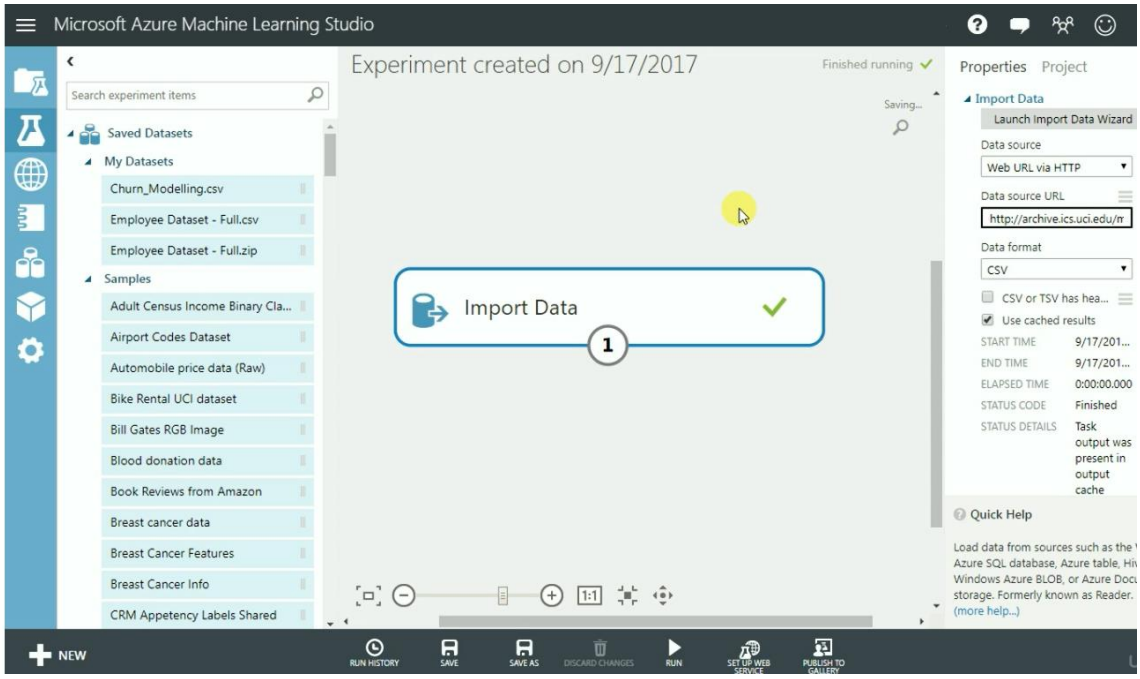
Data found in URL



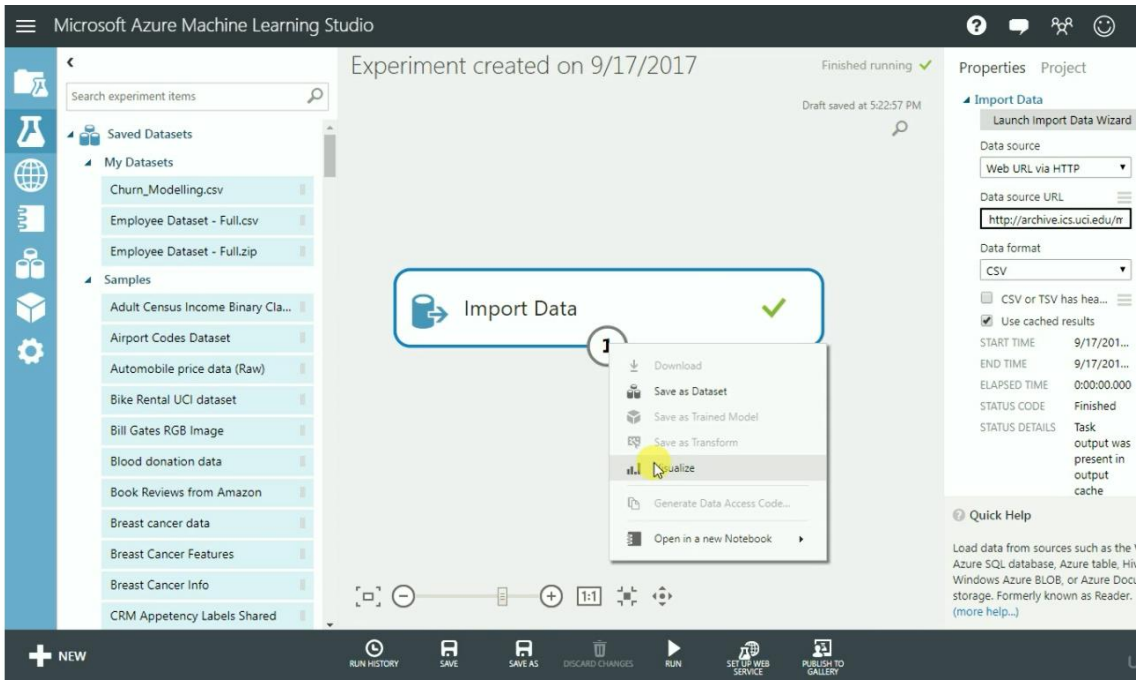
Run the module for result



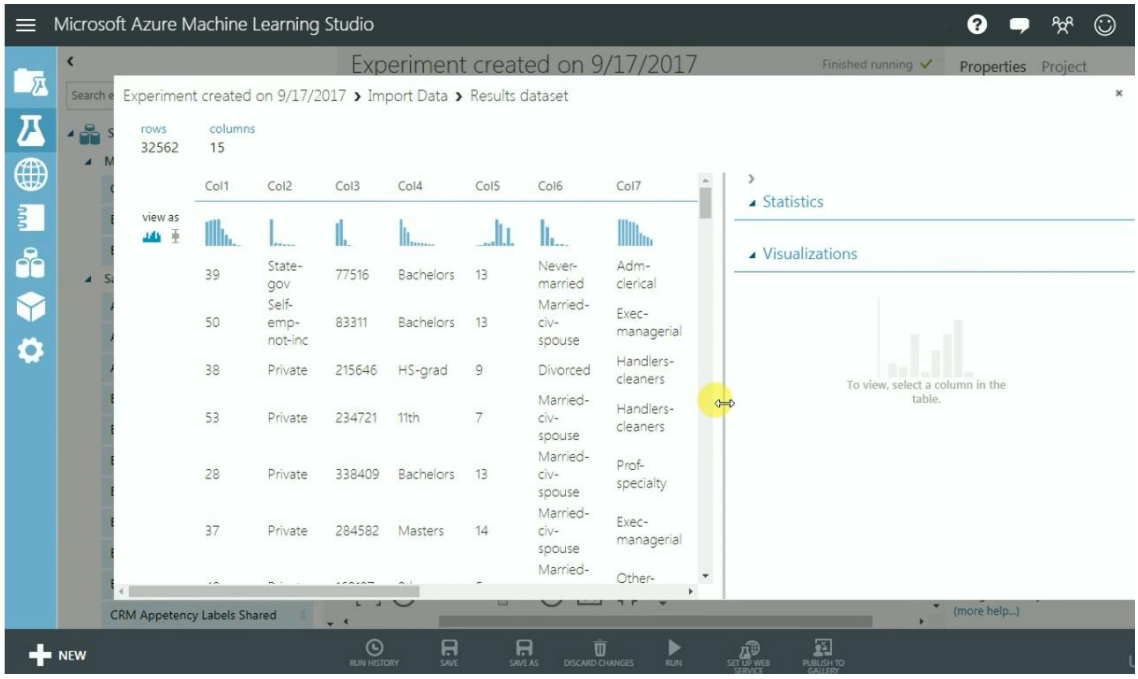
After successful execution visualize the data



Right click the node and visualize the data

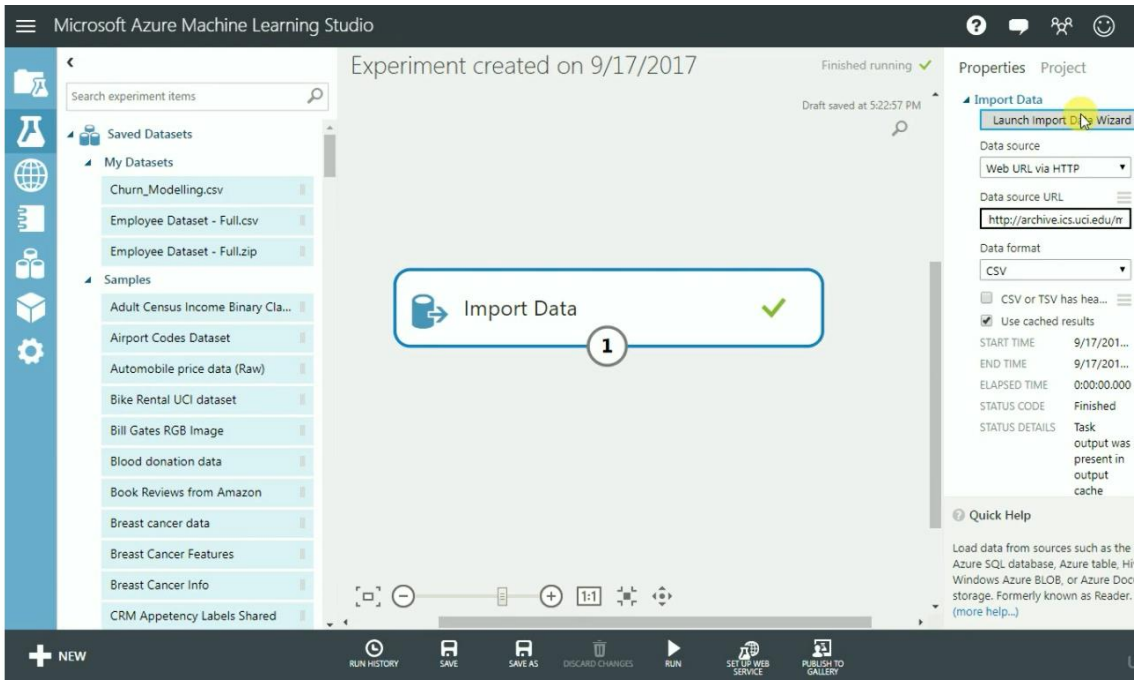


Data imported from URL

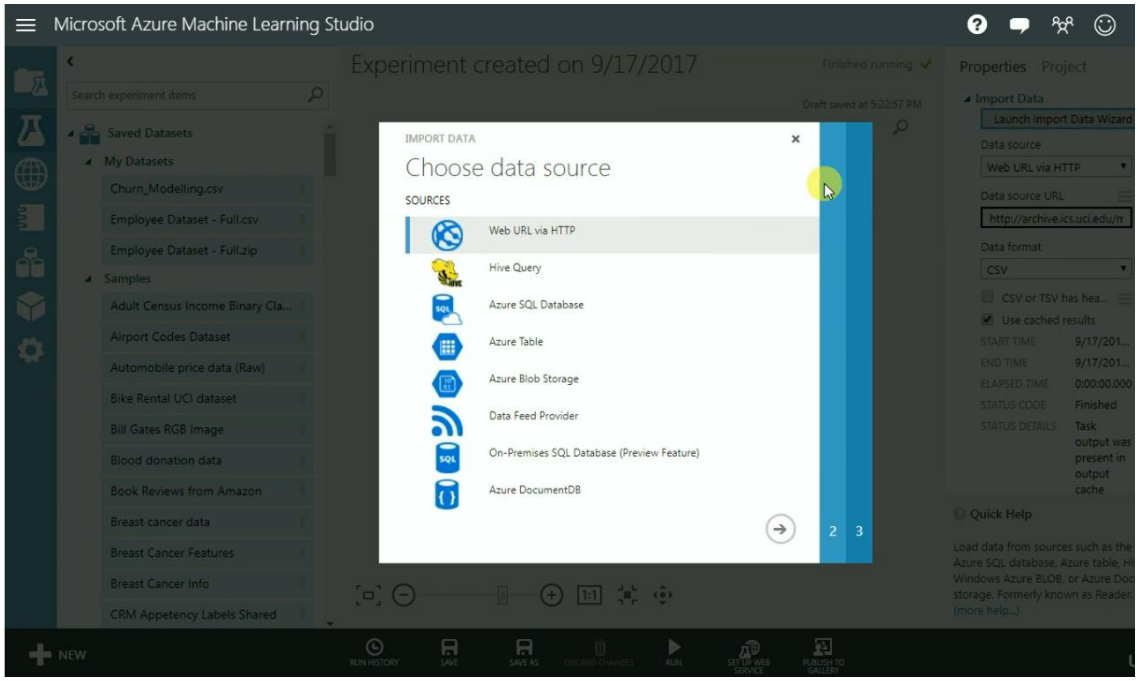


Launch import data wizard

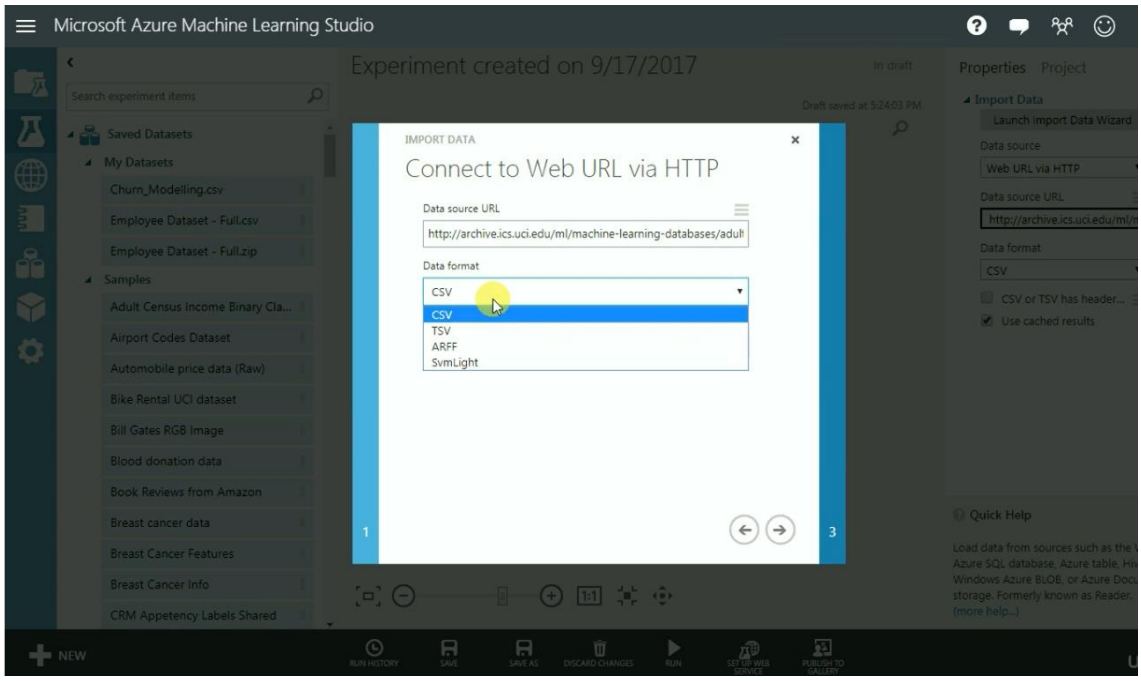
Click Launch import data wizard in the right side



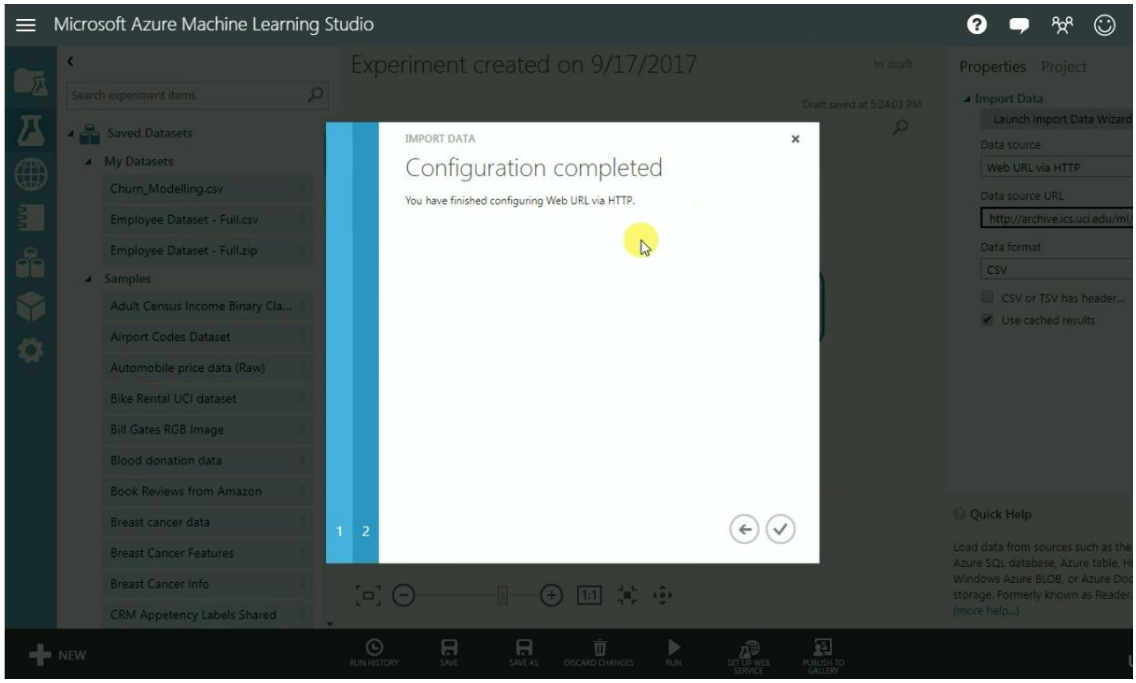
Select web URL from HTTP from the list and click on forward arrow



Enter the data source URL and data format

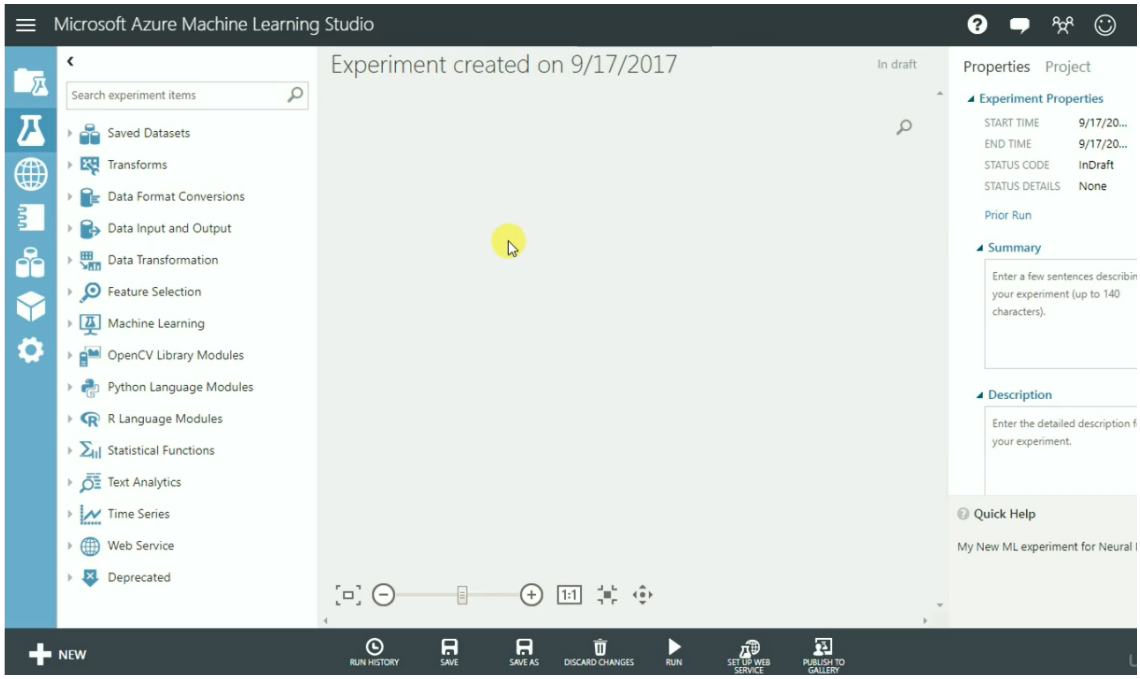


After successful configuration, you can run the module and get same result

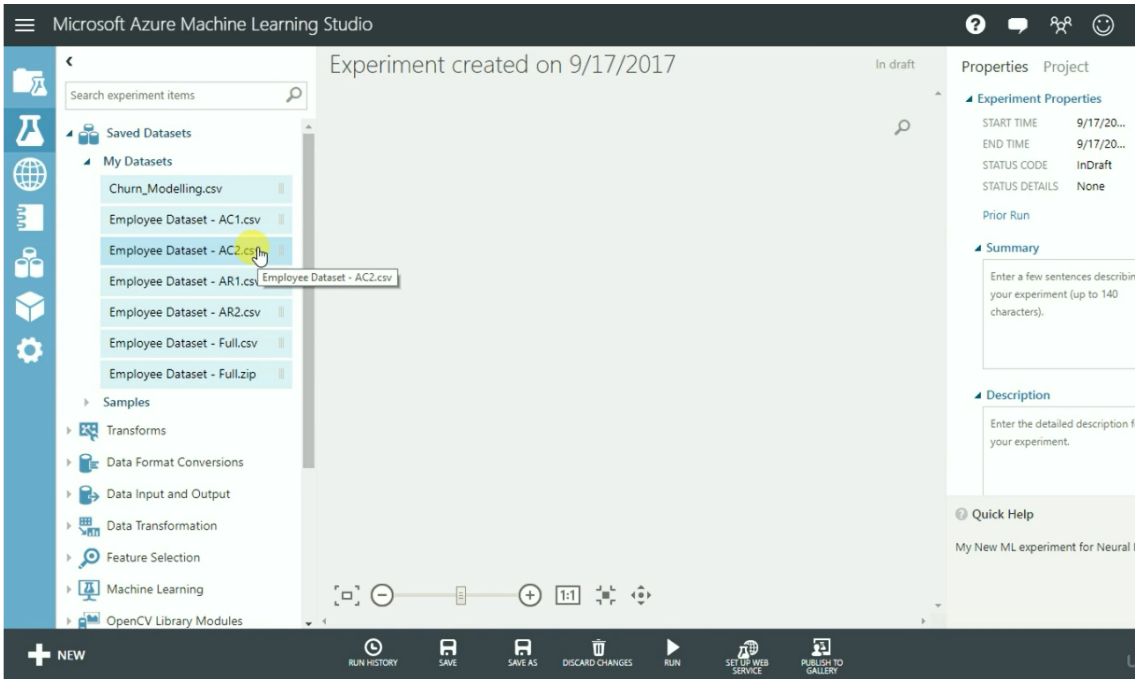


Data Manipulation Using Add Columns Component

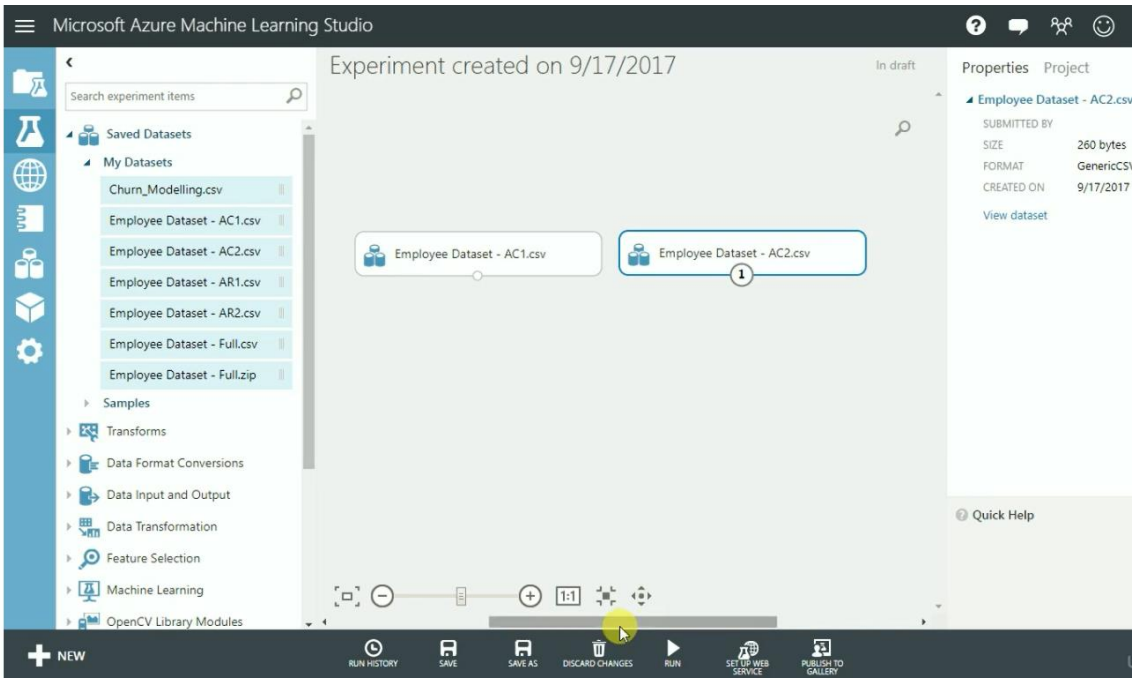
Add Columns



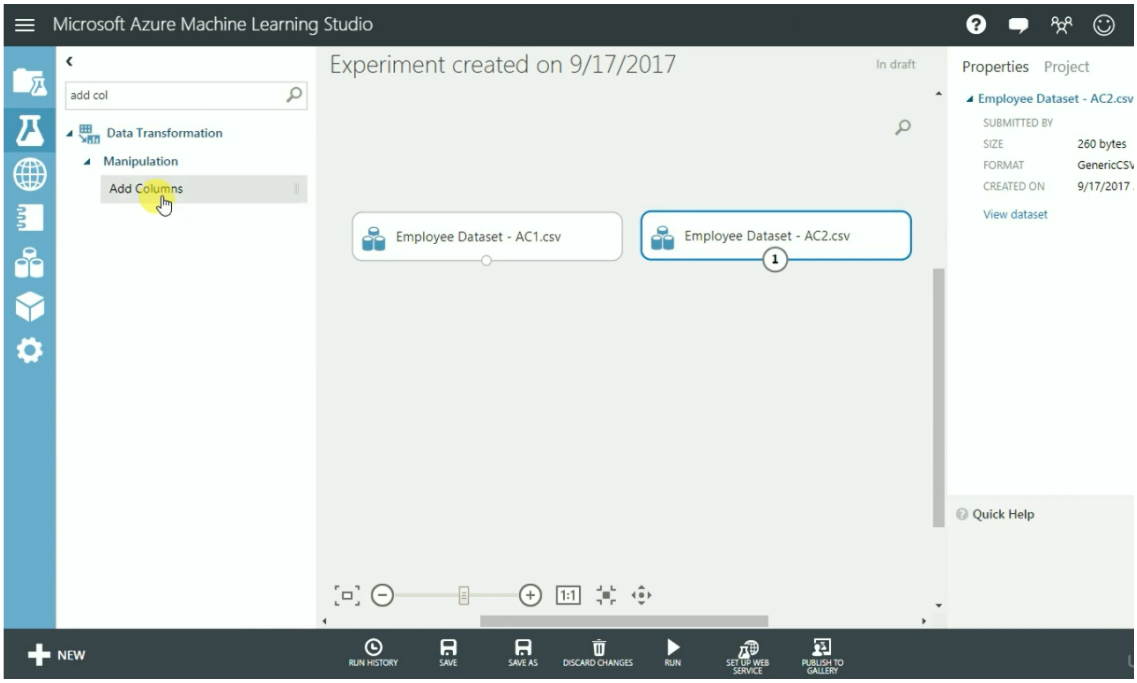
Select any two datasets



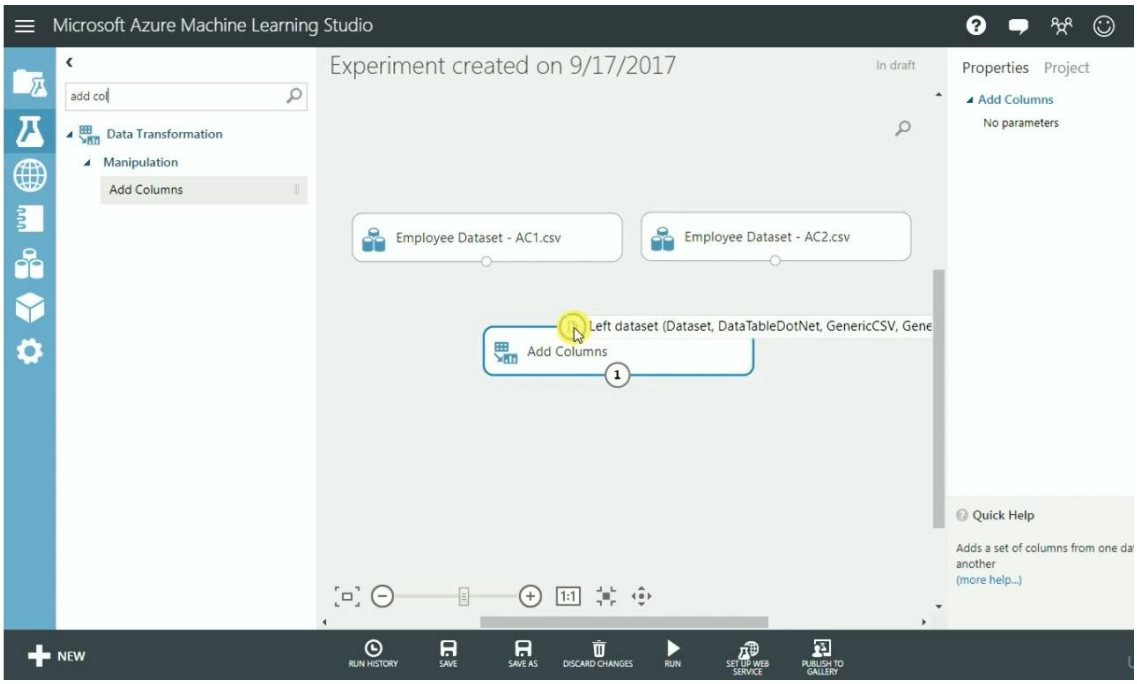
Drag and drop two datasets from the list



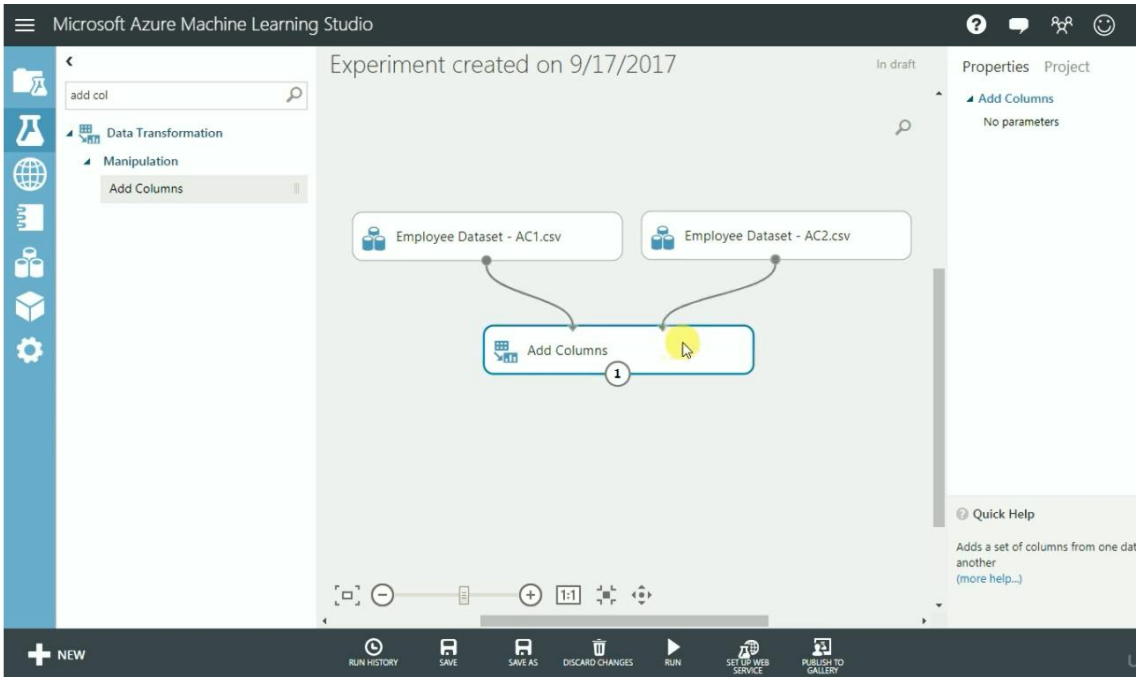
Search for add columns and drop the same in canvas



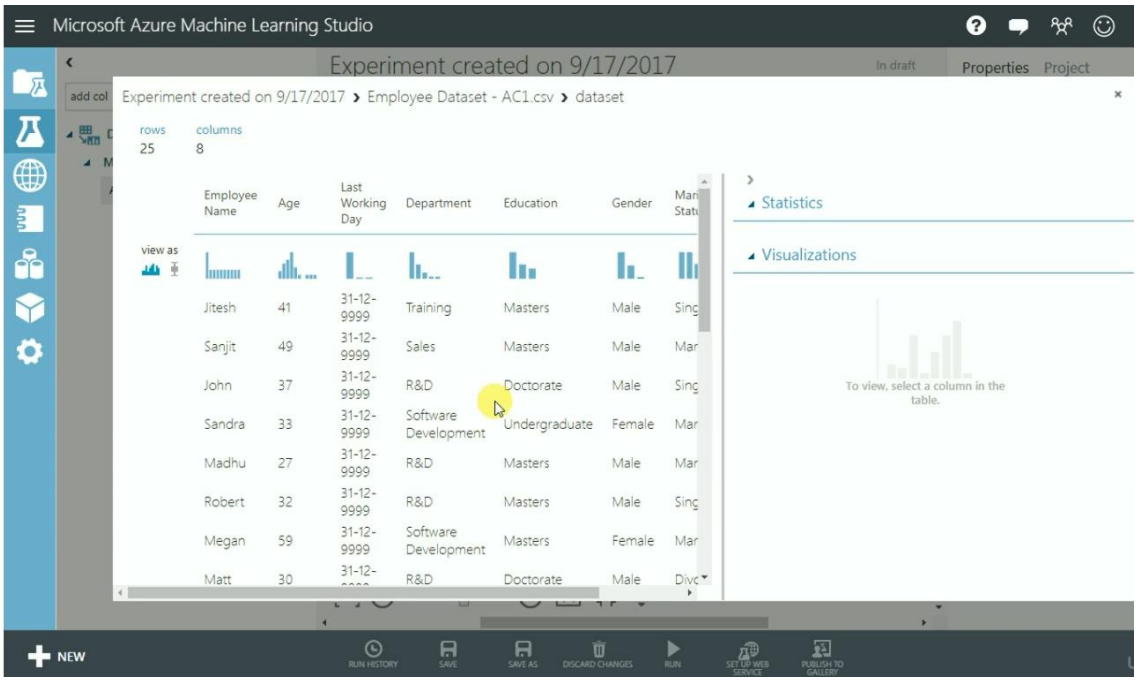
Add column has left dataset and Right dataset

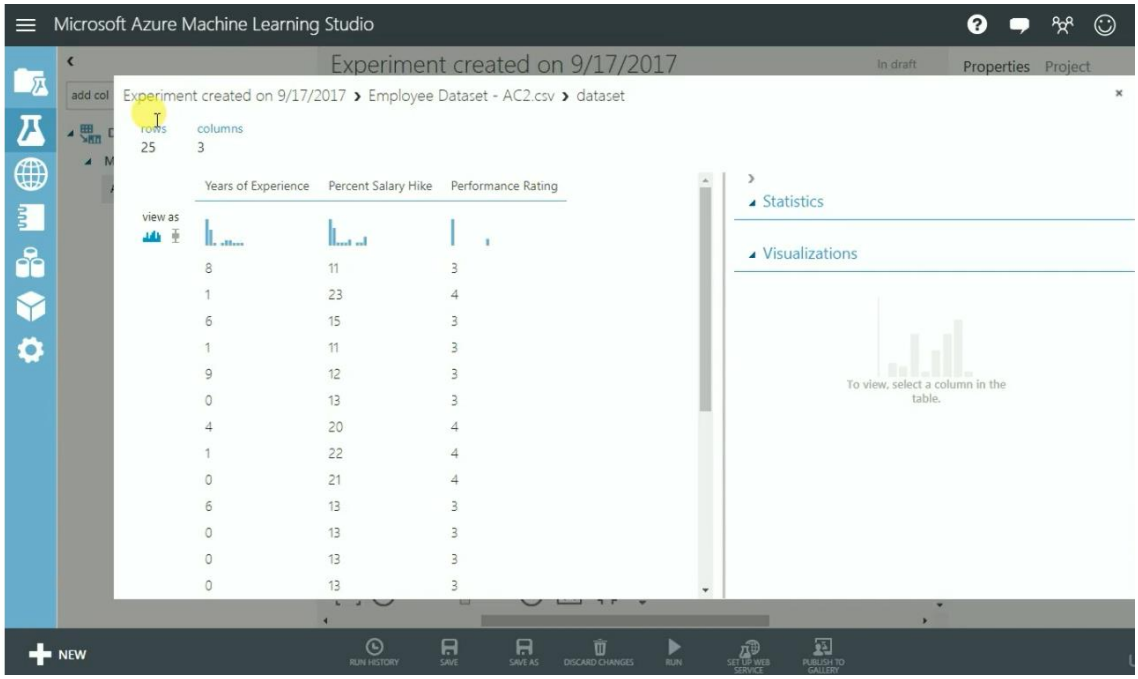


Connect both csv files to Add Column Left and right input node

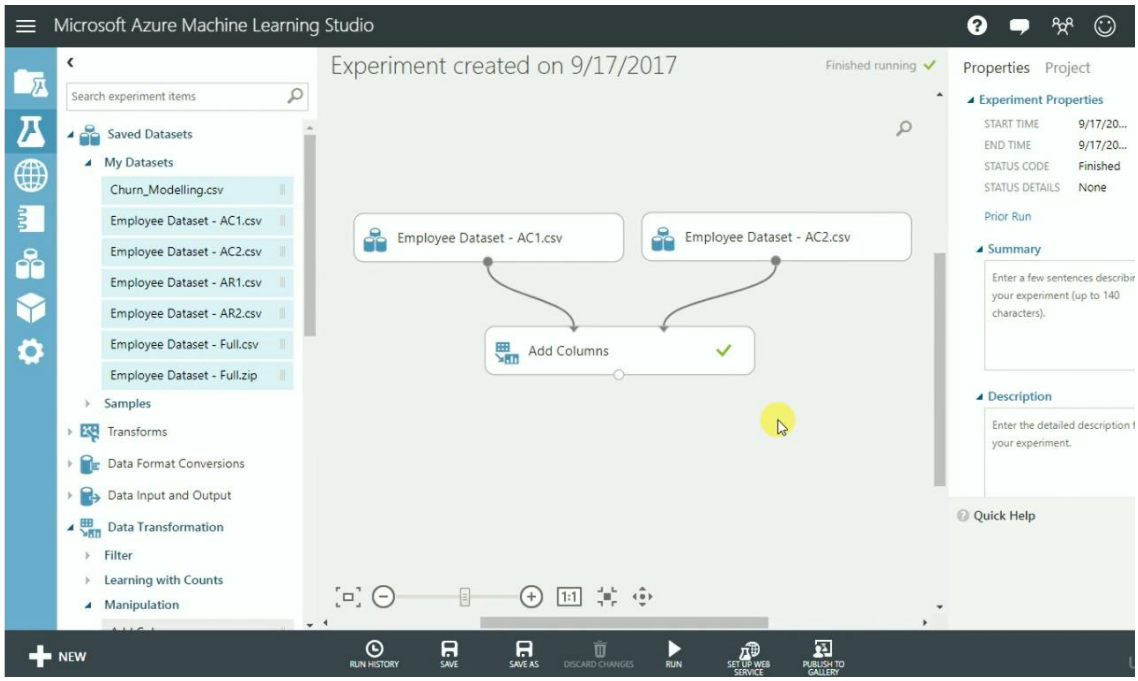


Visualize both csv files before execution to know its columns

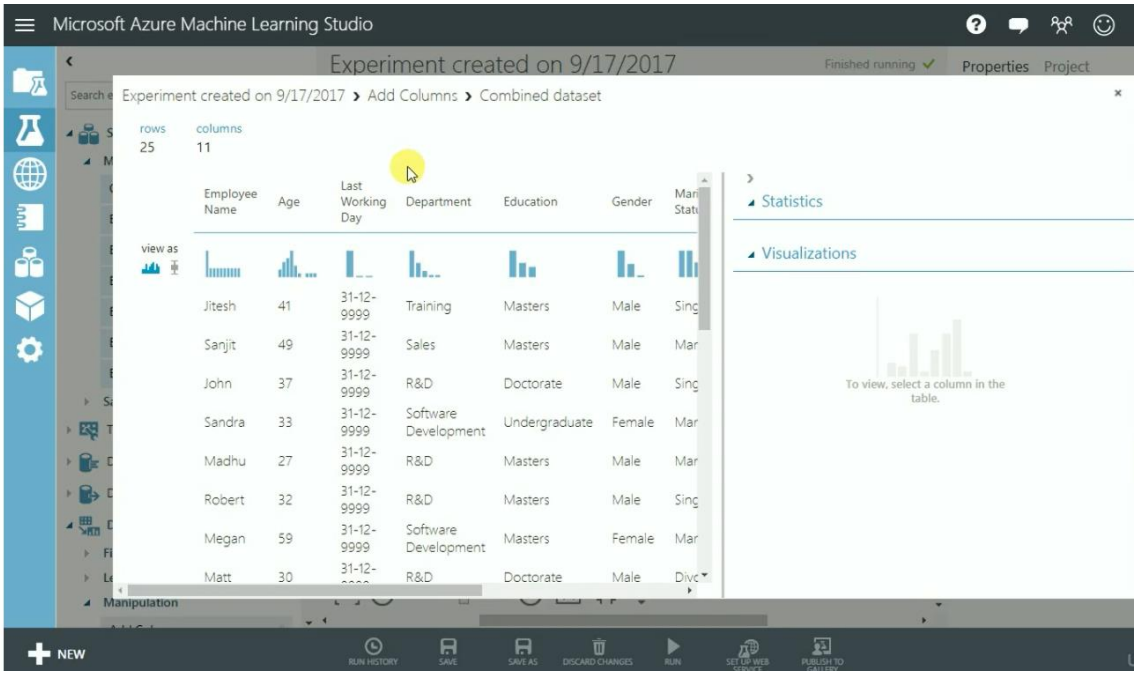




Run the module and click visualize



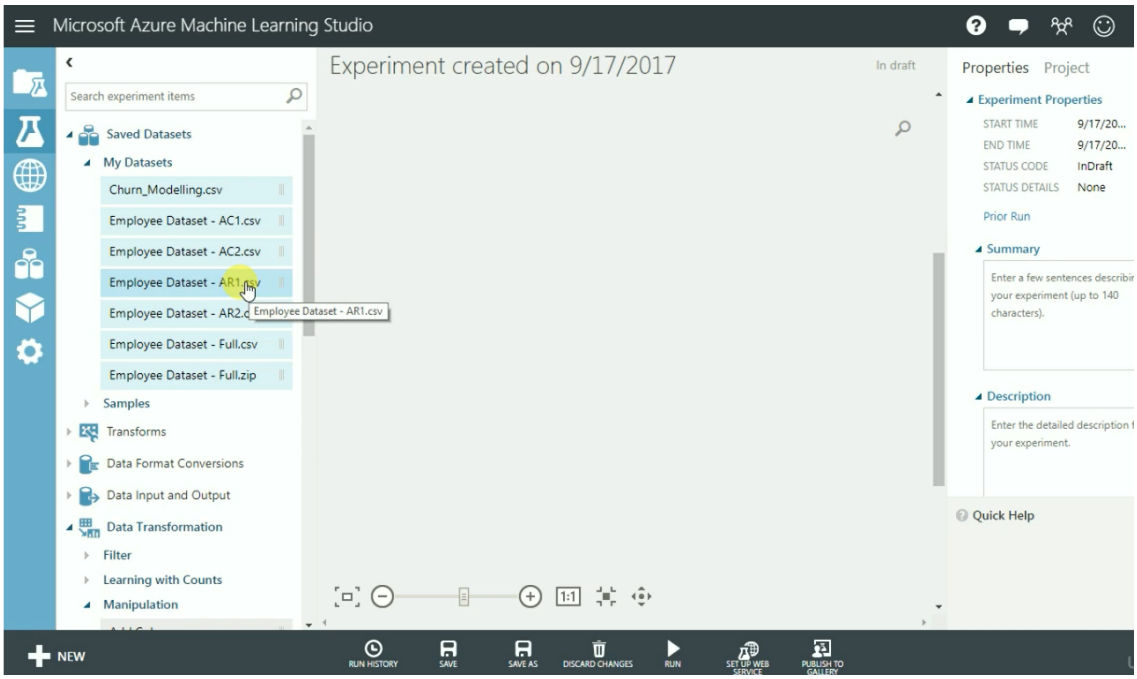
Result is as expected that columns has been updated



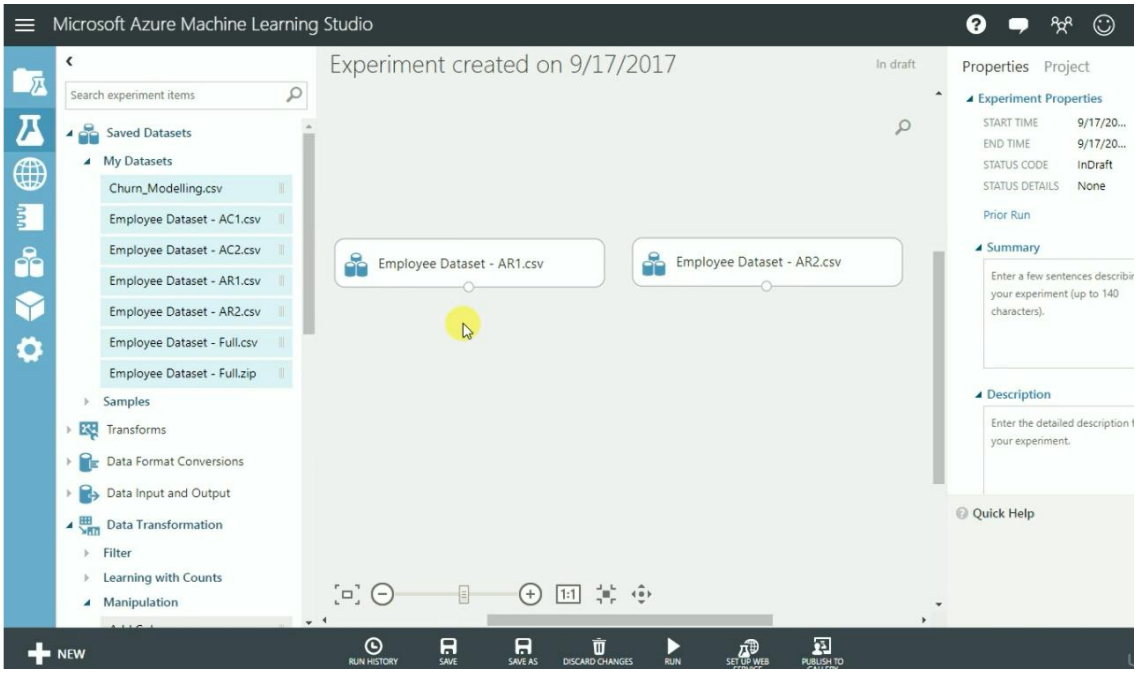
Data Manipulation Using Add Rows Component

Add Rows

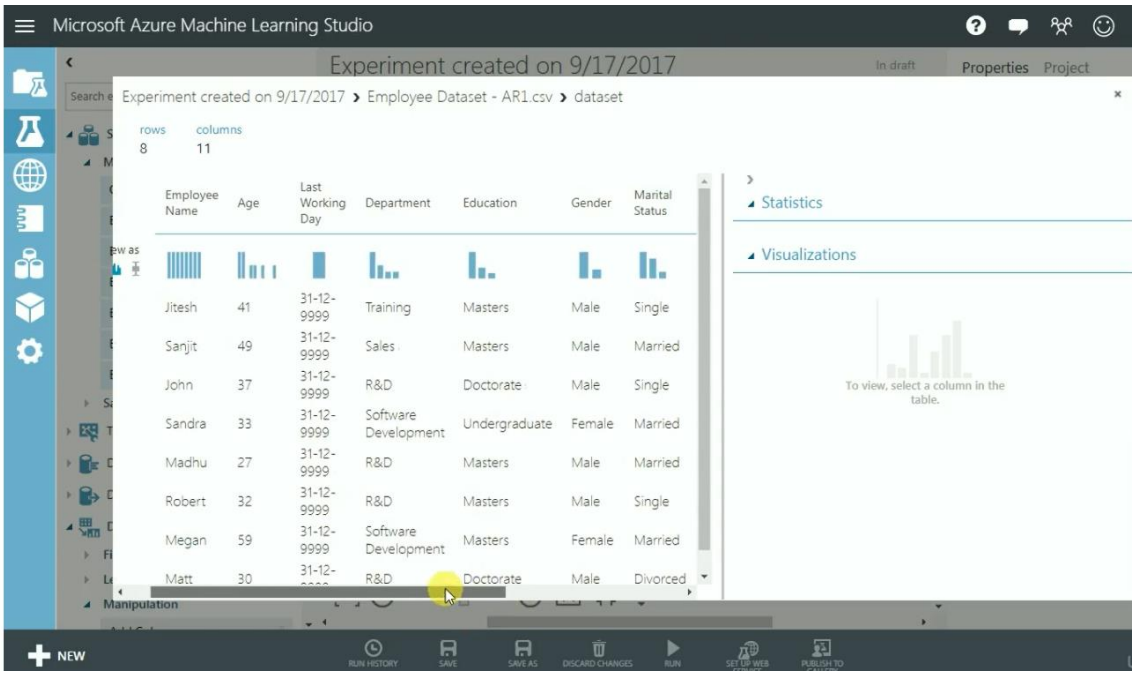
Select datasets for Adding rows

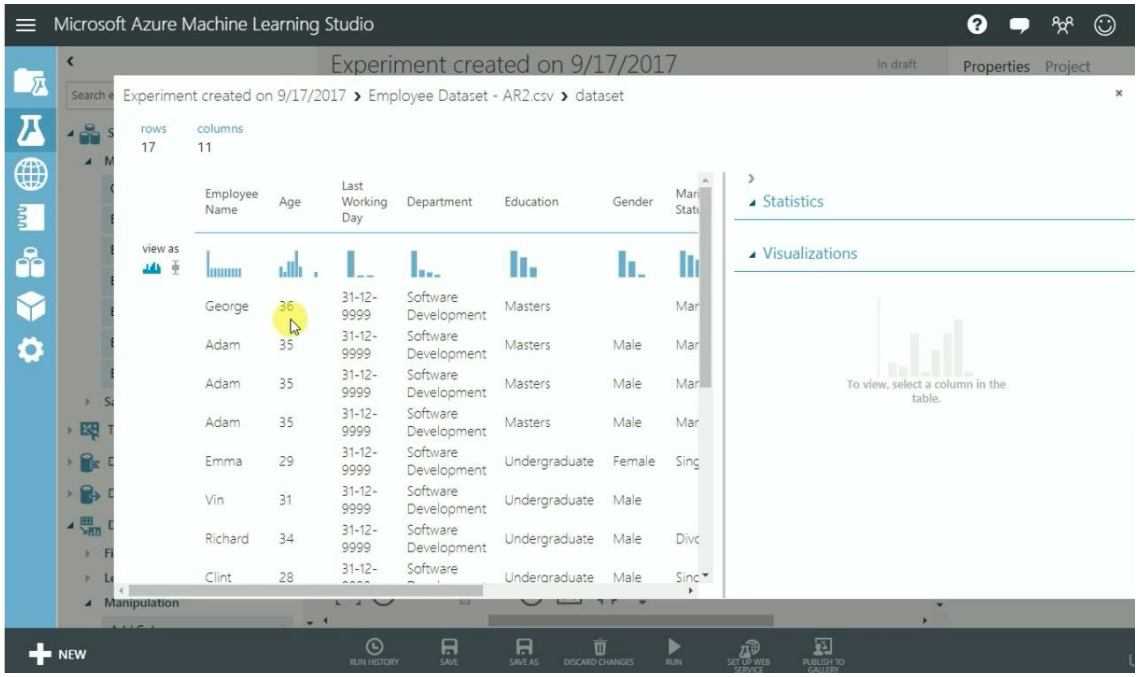


Drag and drop selected datasets

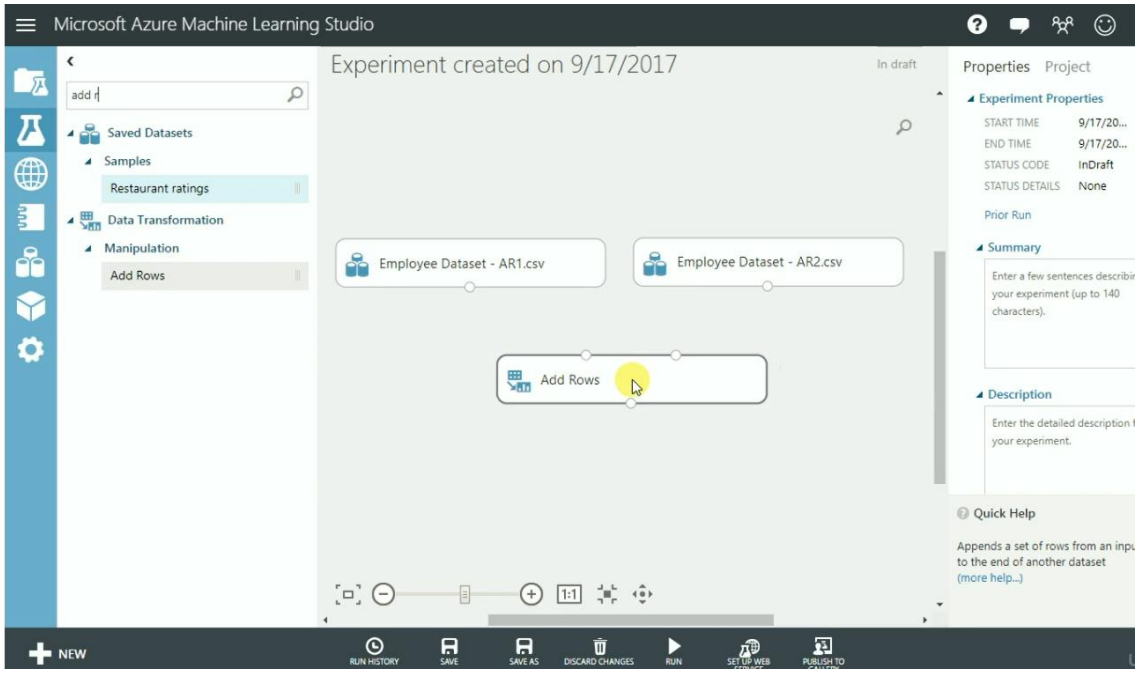


Visualize both csv before executing

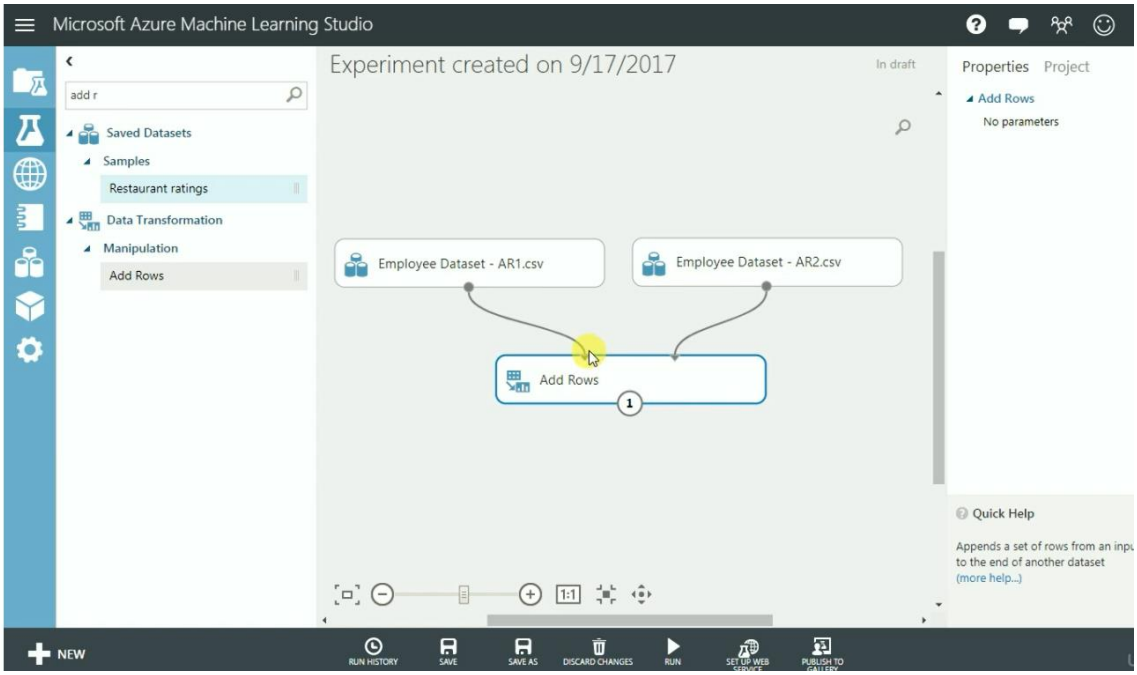




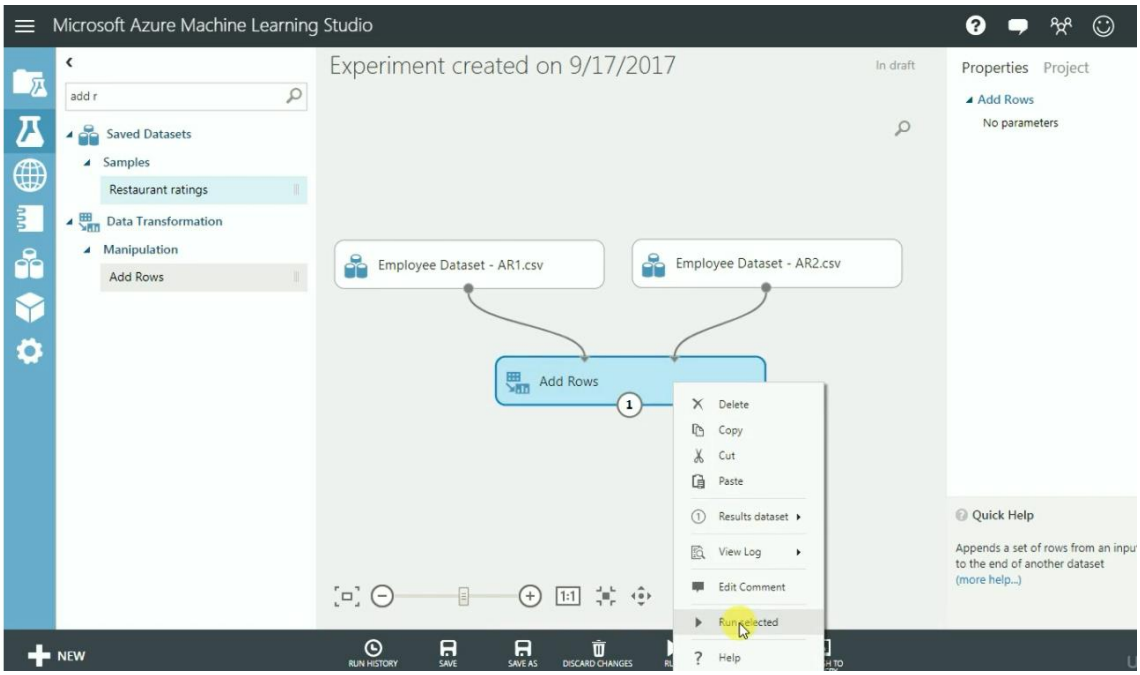
Search for add rows, drag and drop in the canvas



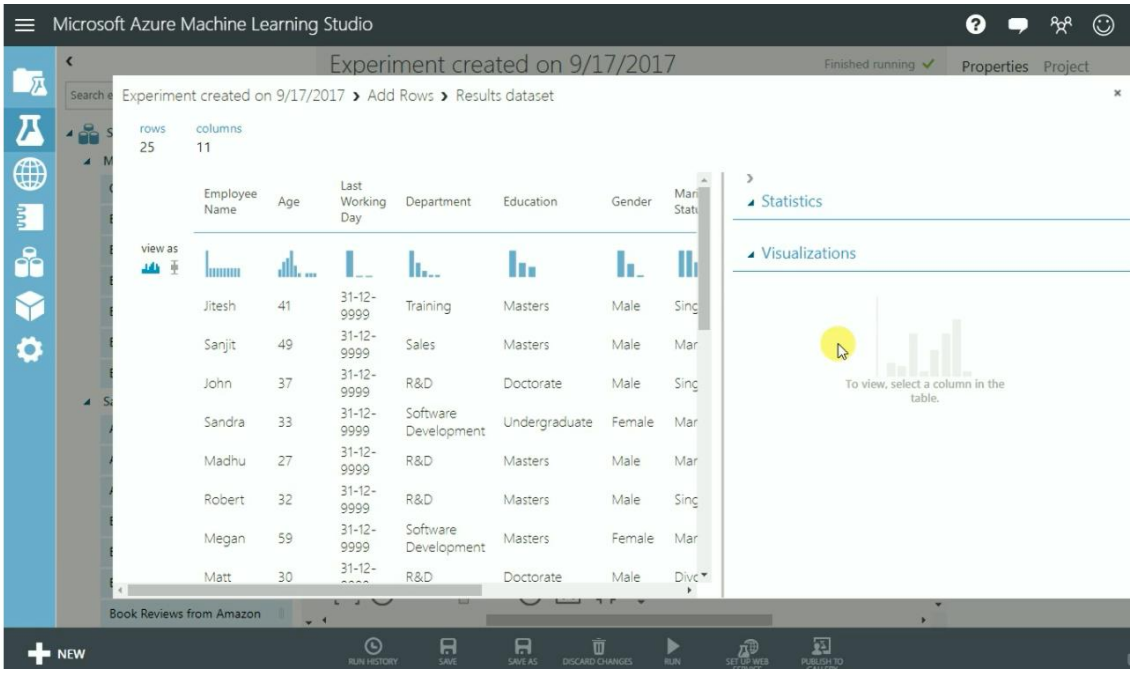
Connect both the output nodes from csv to Add row



Run and execute



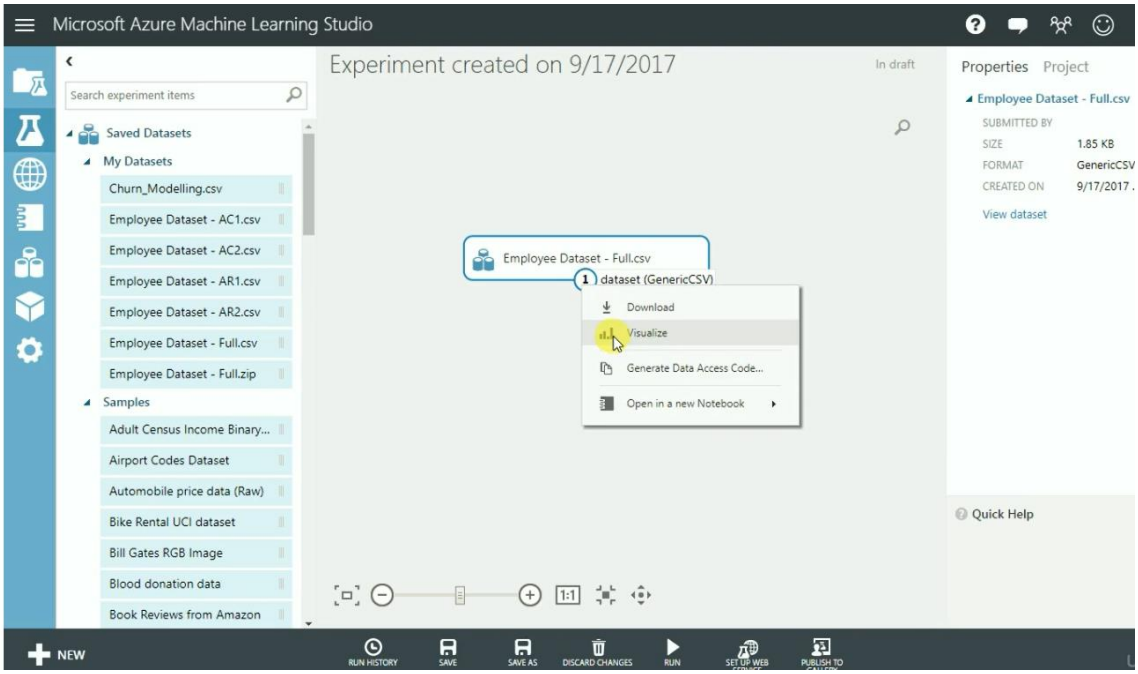
Run and visualize the output



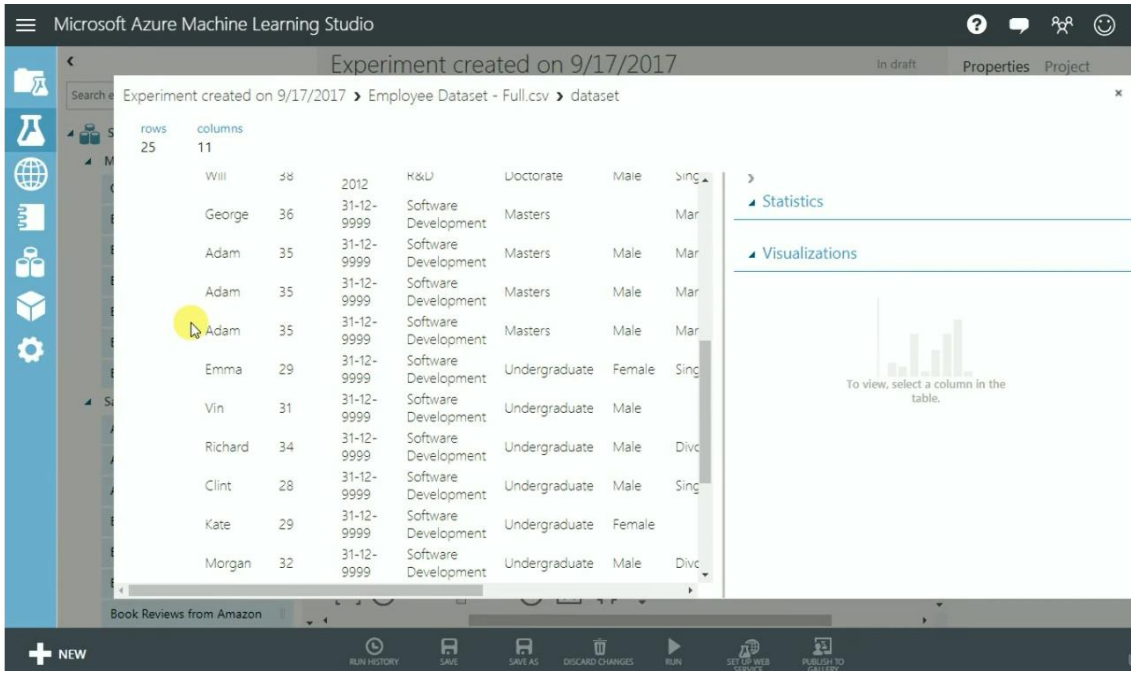
Data Manipulation Using Remove Duplicate Components

Remove Duplicate

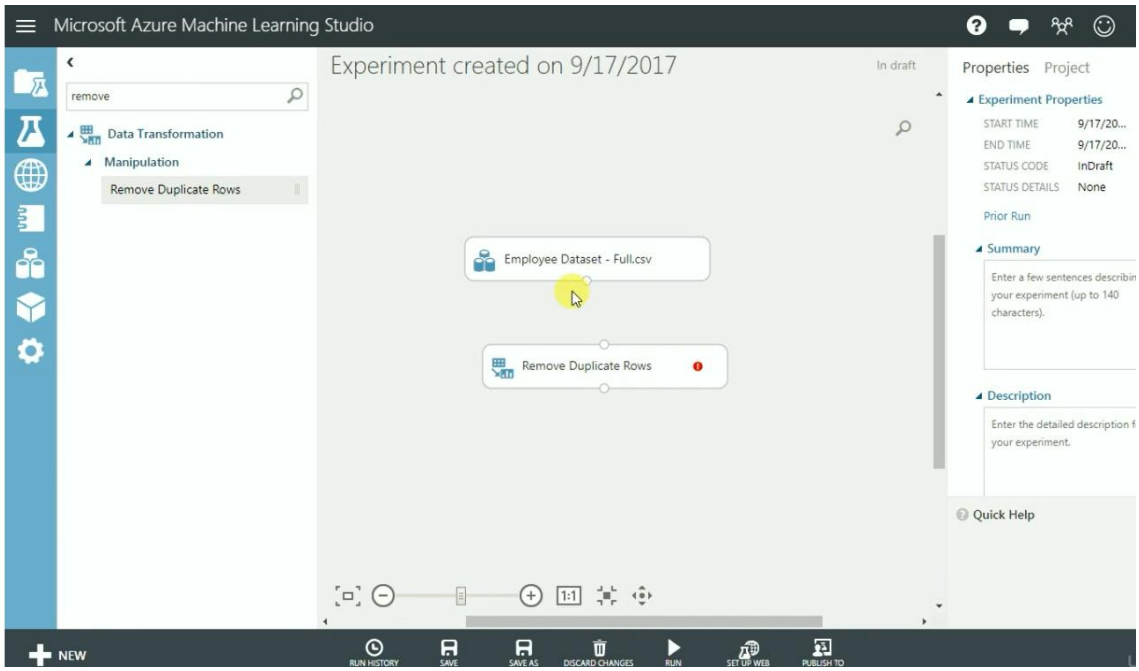
Visualize a data set with duplication



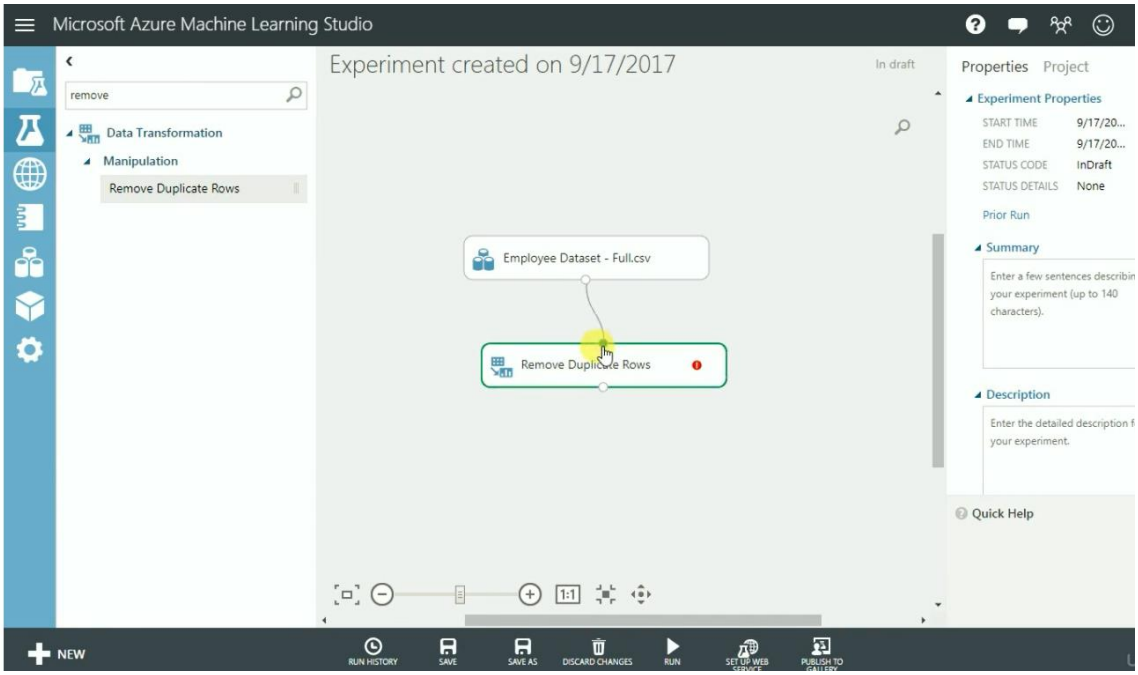
Can see three duplications in the visualization



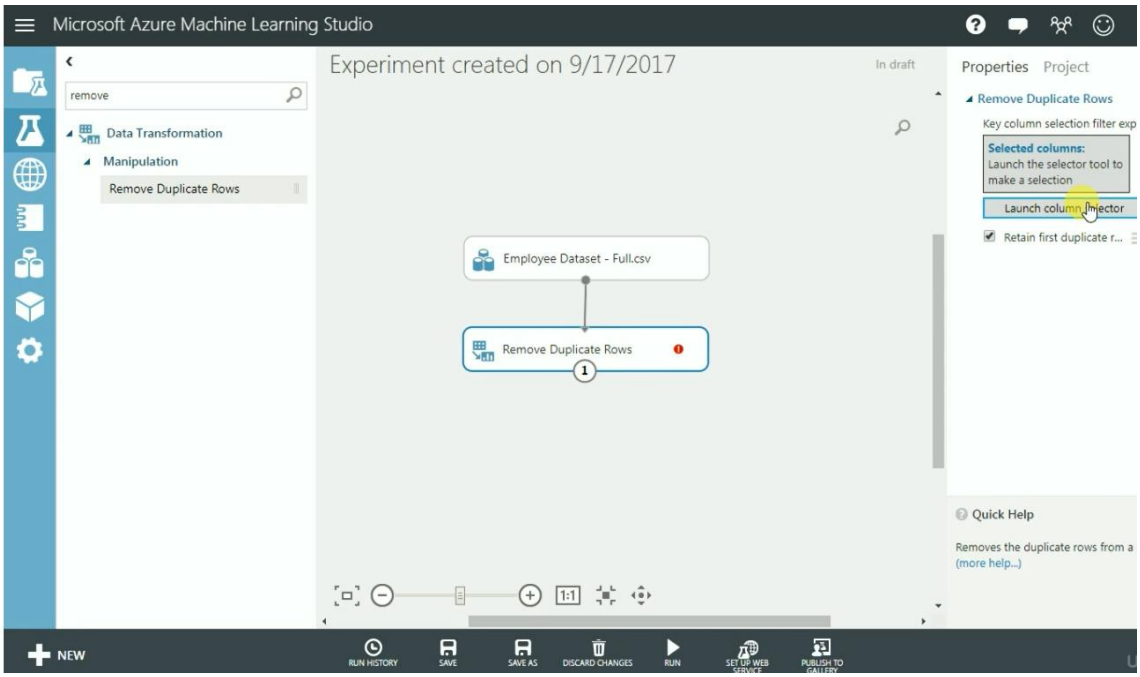
Search for remove duplicate rows, drag and drop the same in canvas



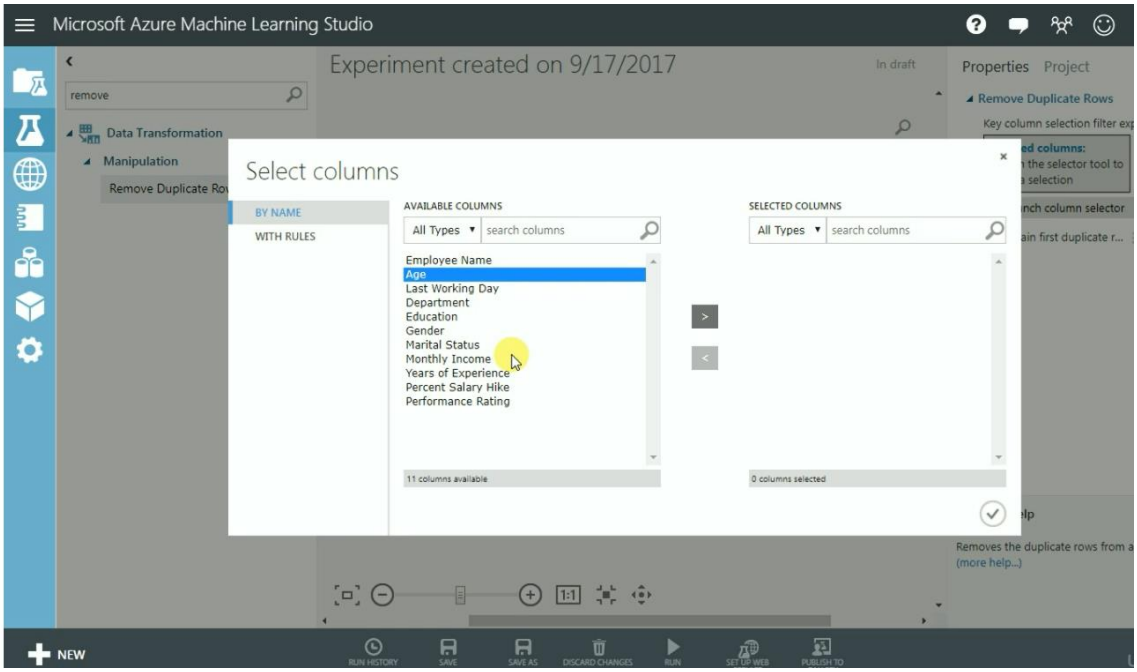
Connect both input and output nodes



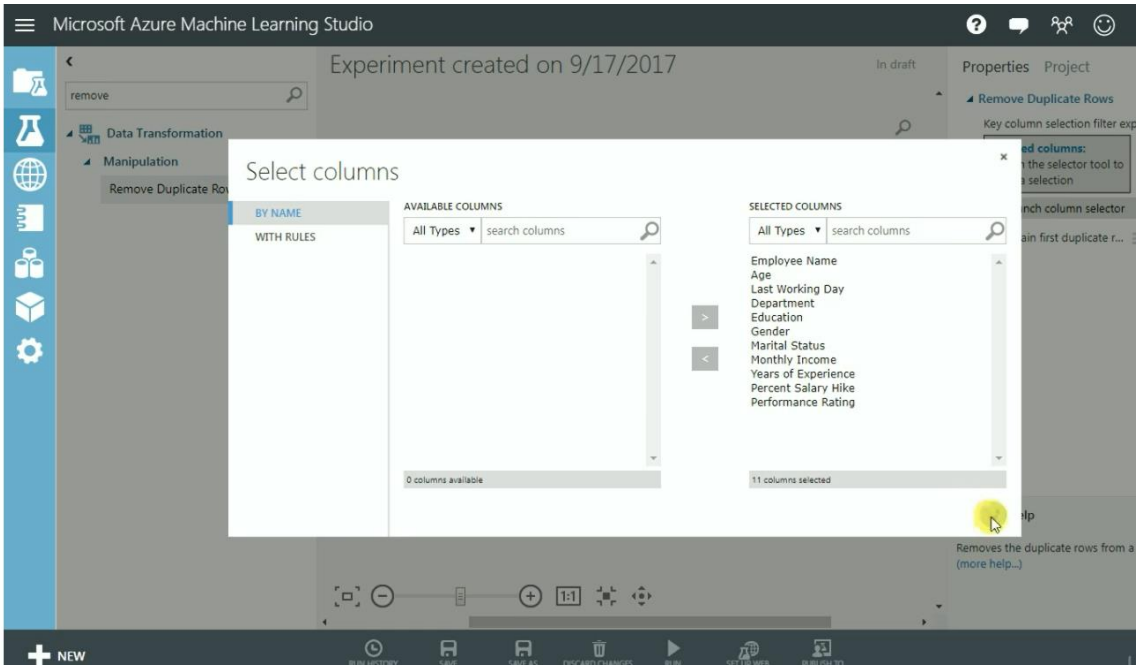
Click on Launch column selector



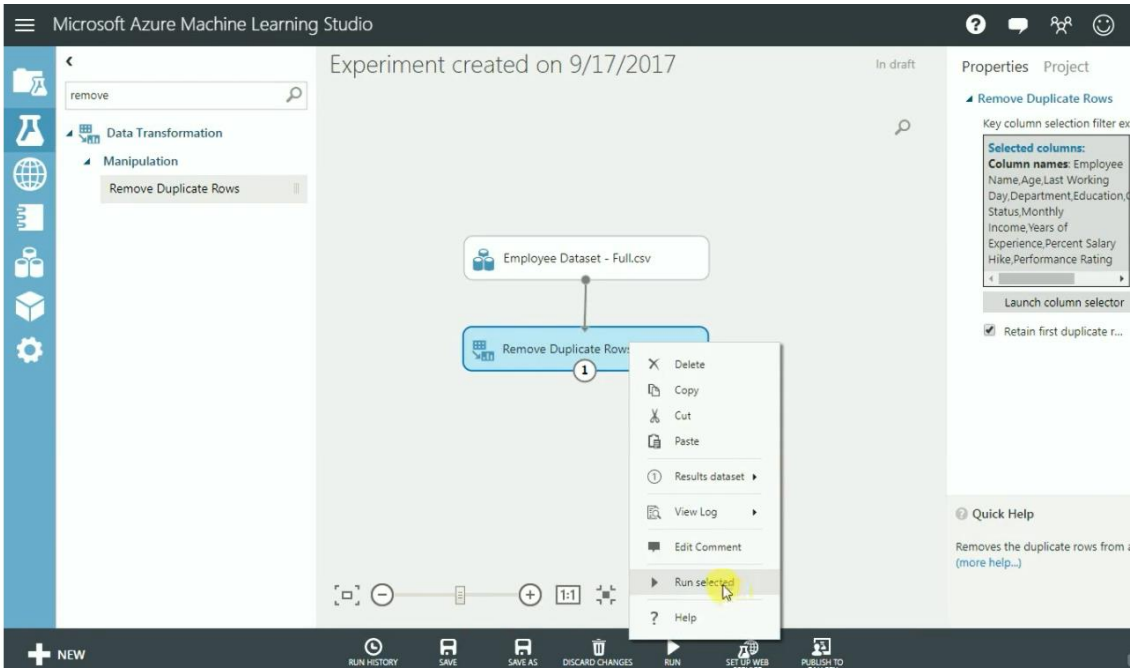
From the available columns select all the types and move to selected columns



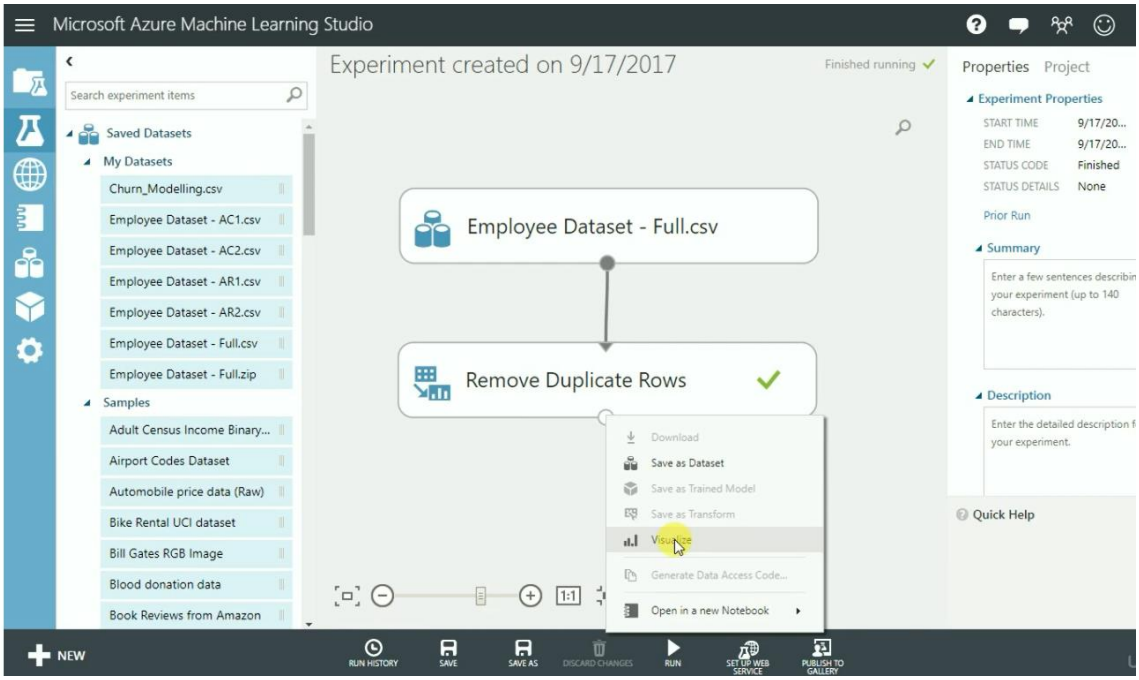
Select all columns and click ok



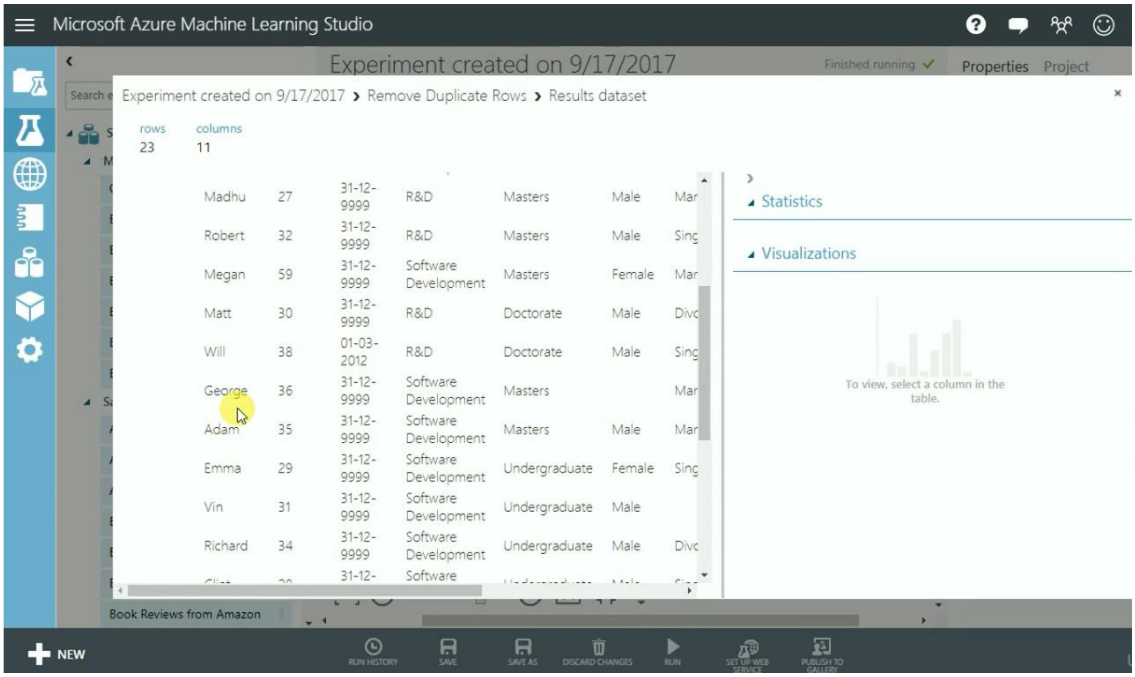
Run the module



Visualize after successful execution

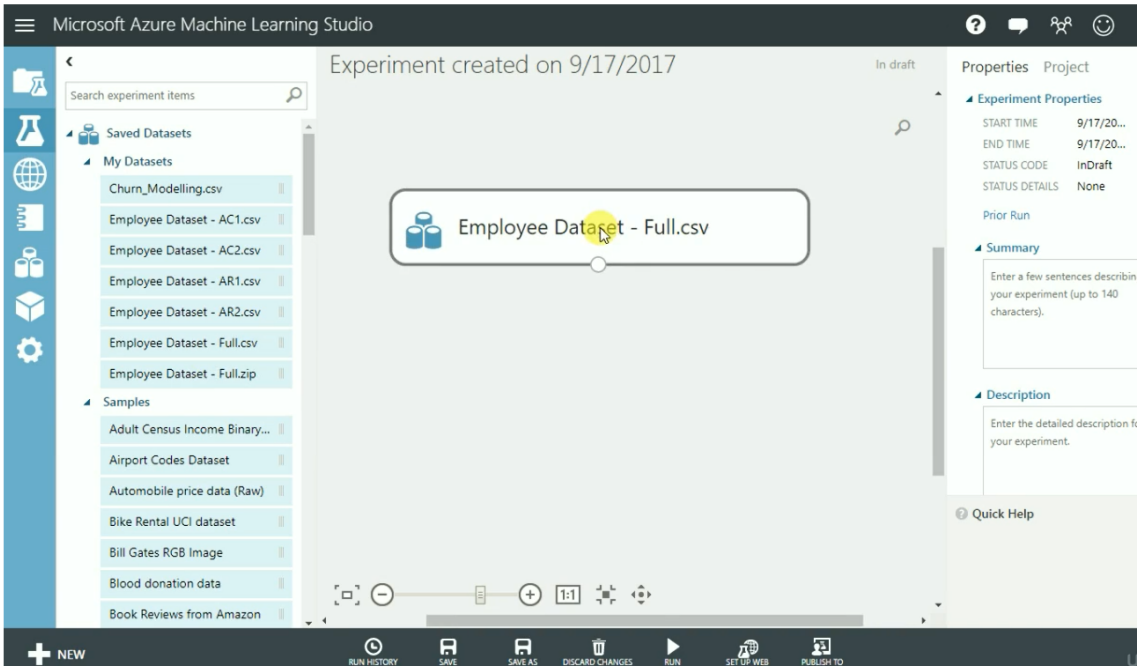


Can view that duplication removed successfully

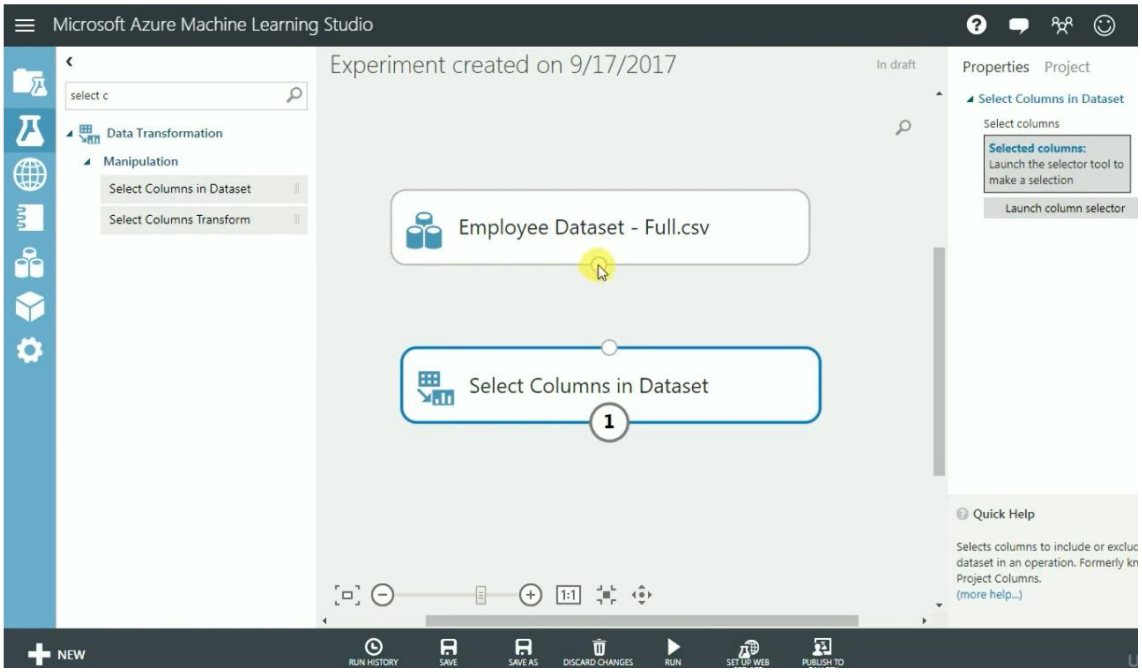


Data Manipulation Using Select Column in a Dataset Component

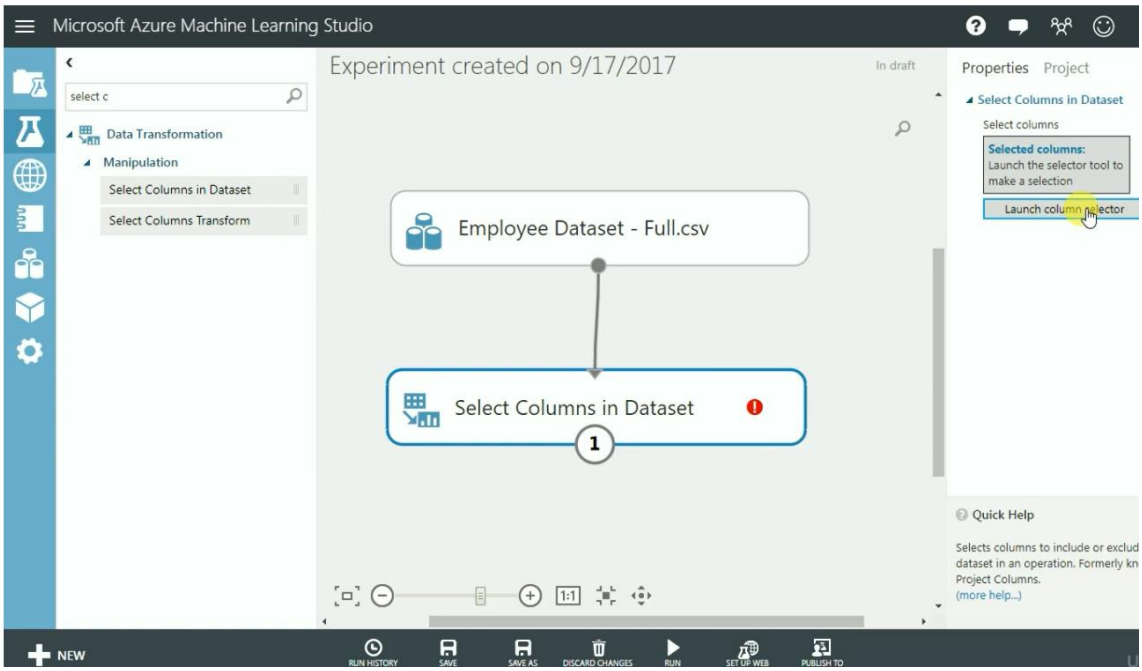
Take a dataset which is uploaded already in canvas



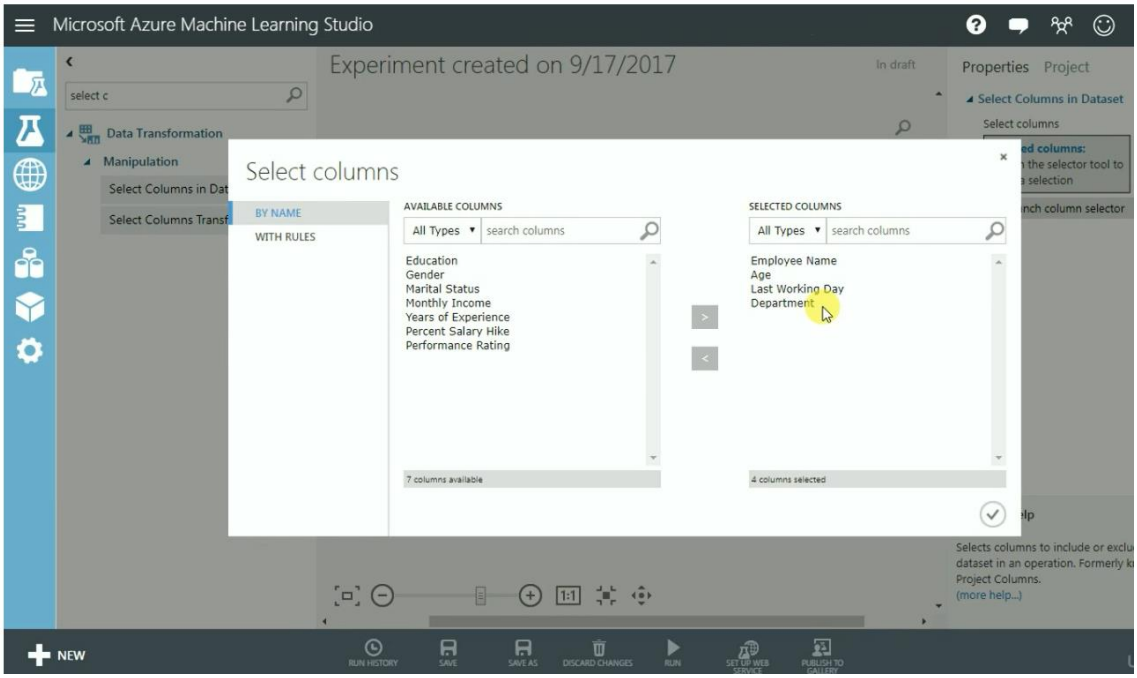
Search for select columns for datasets, drag and drop in canvas



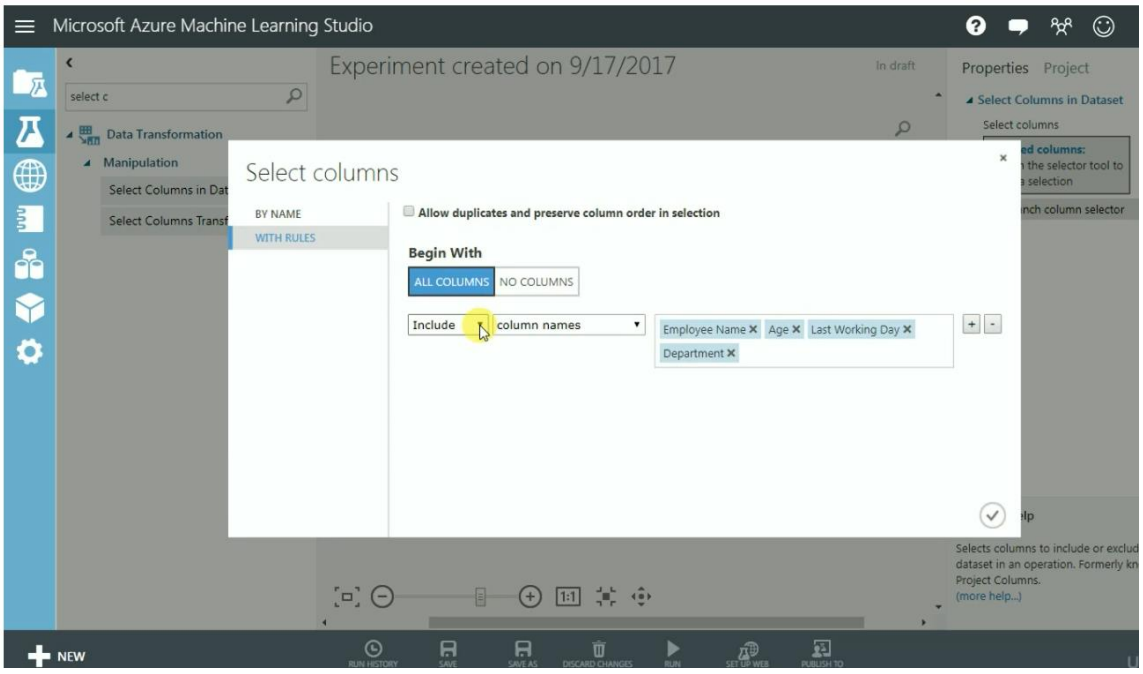
Connect both output and input nodes and click on launch column sector



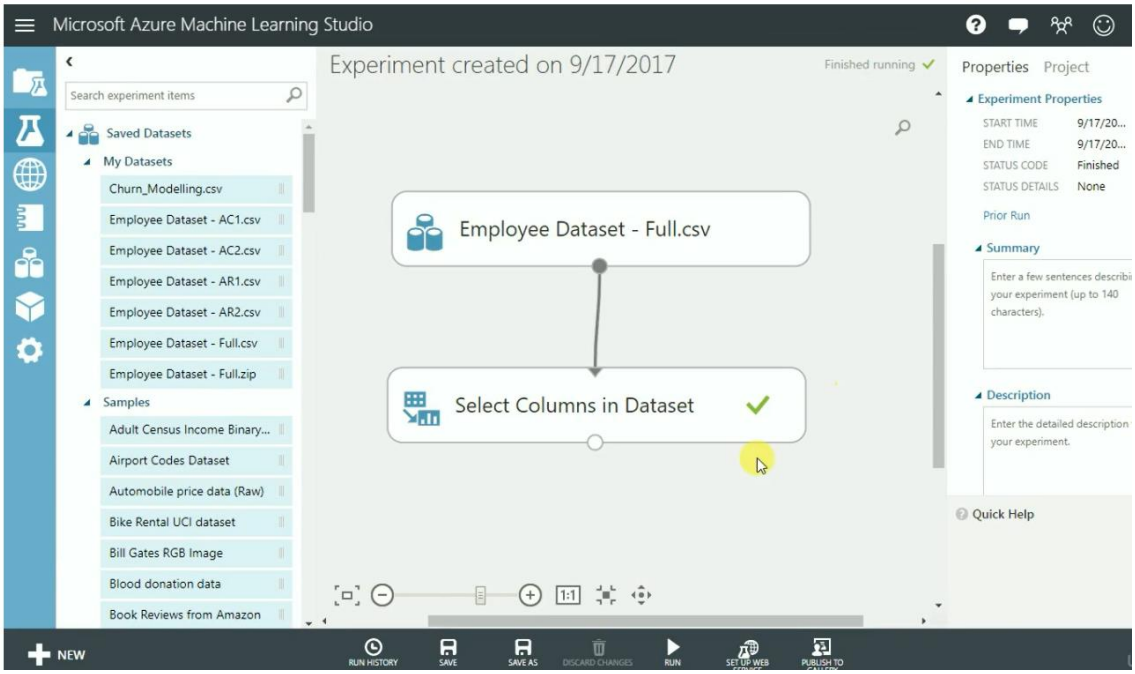
Select first four types and move to selected columns as below and press ok



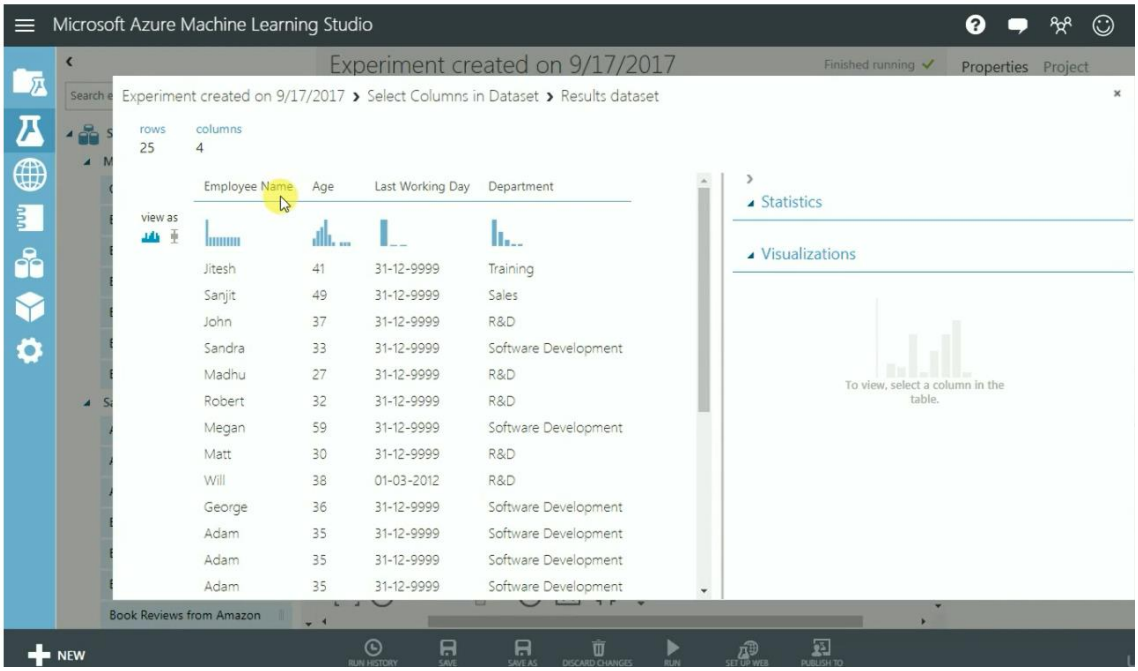
By clicking (with rules) can select any types as per requirement



Run the module after selection and visualize the output

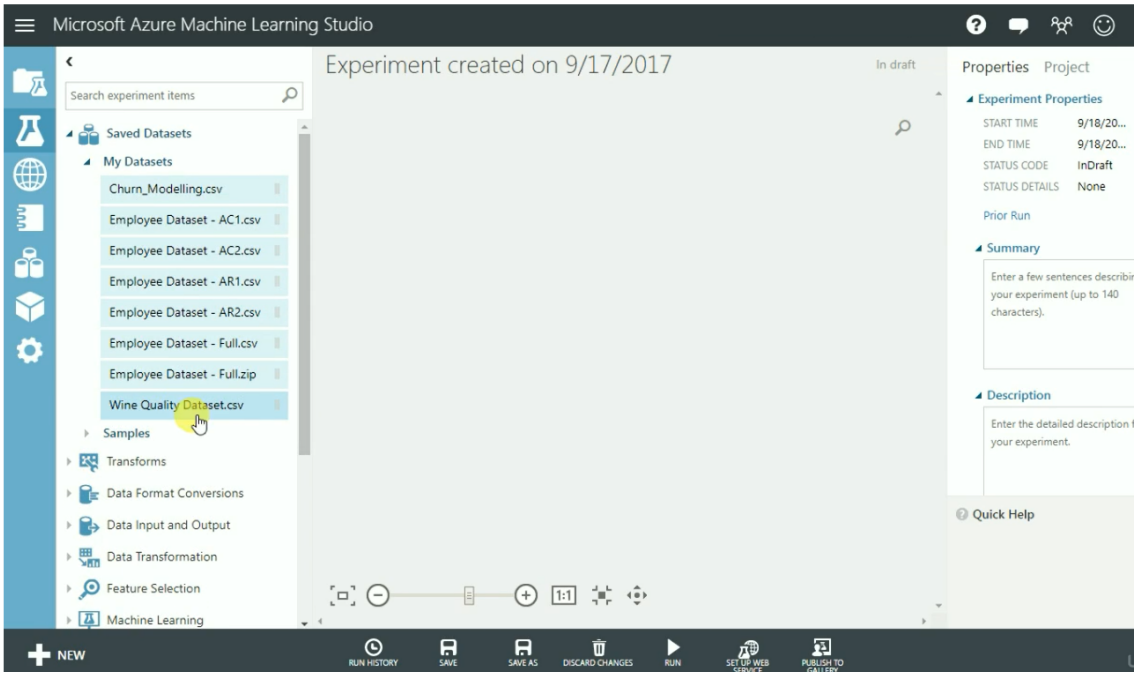


As per requirement output obtained for selected four columns successfully

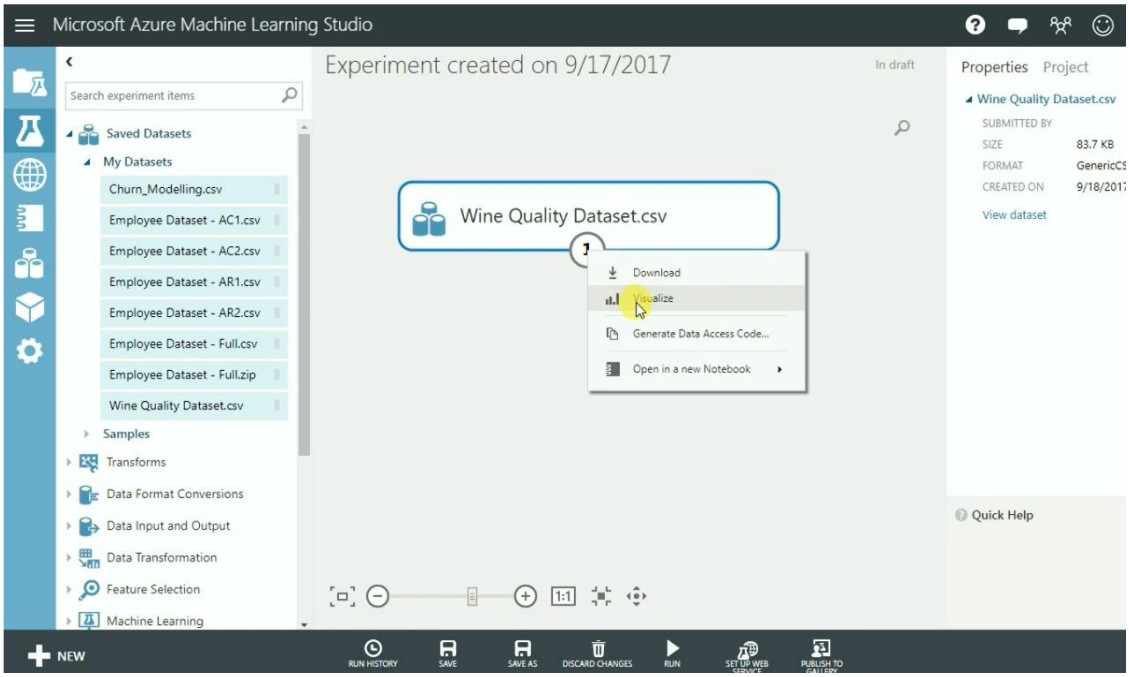


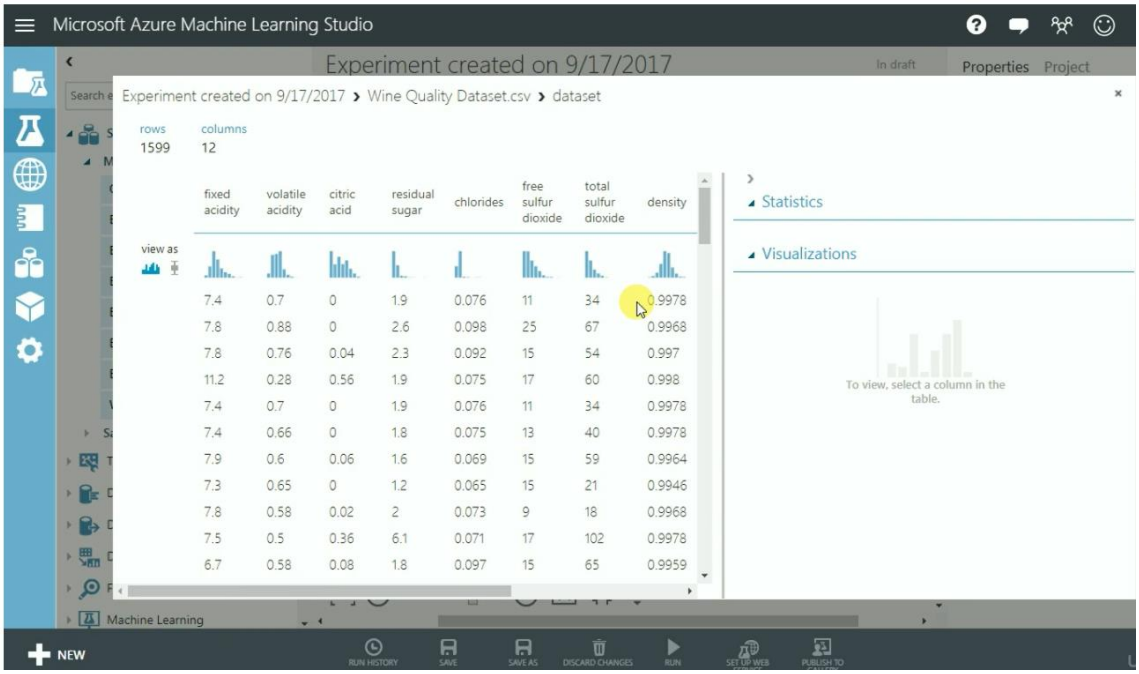
Data Manipulation Using Apply SQL Transformation Component

Select the dataset uploaded

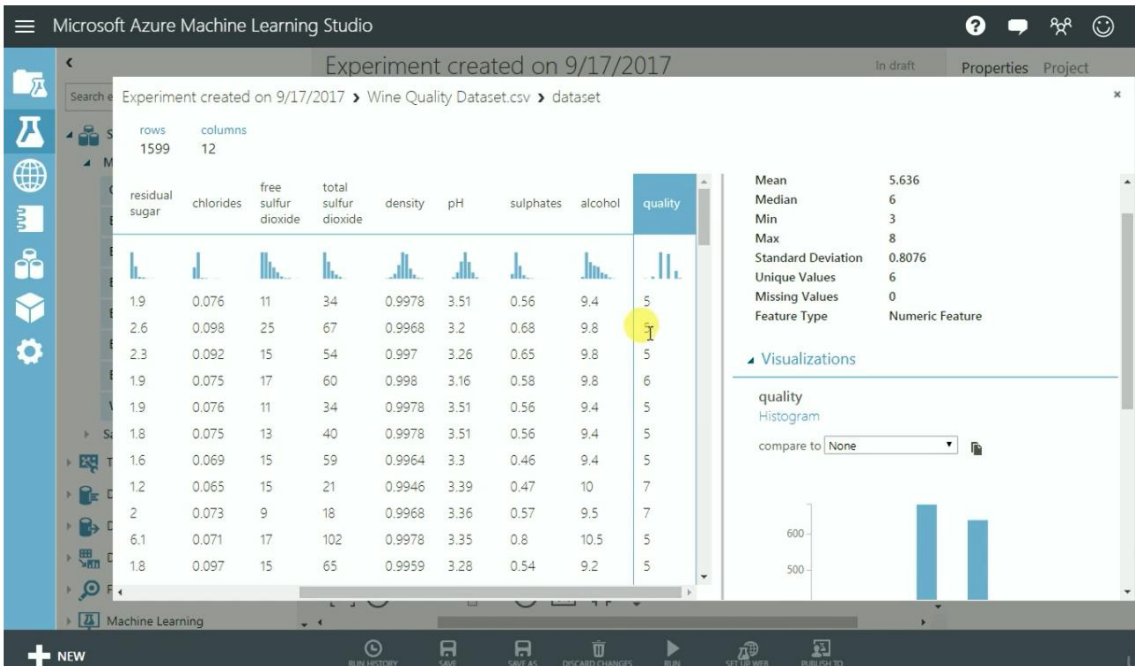


Drag and drop the selected data set and visualize the same

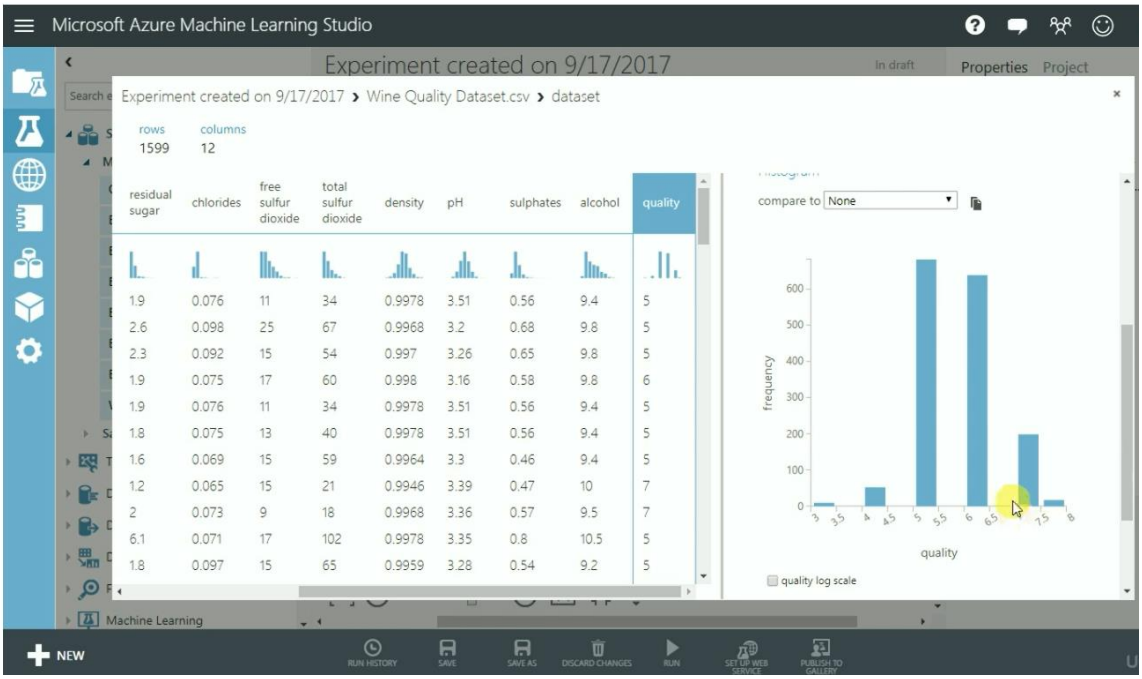




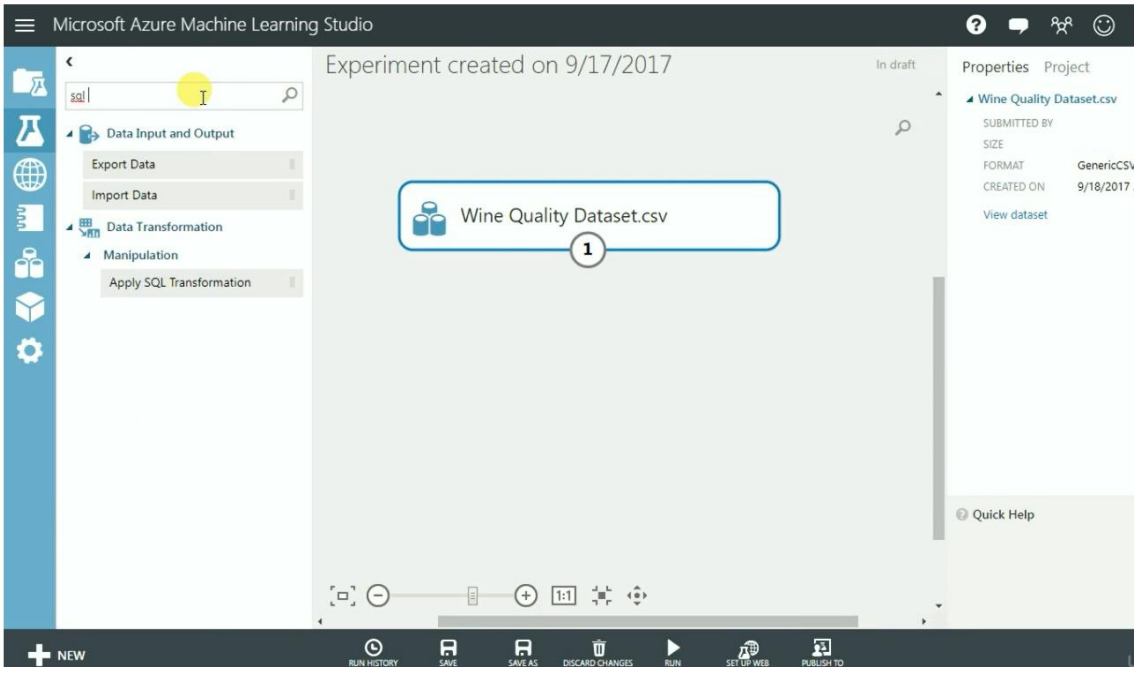
This dataset used for checking quality of wine, we can check the classification High, medium and low



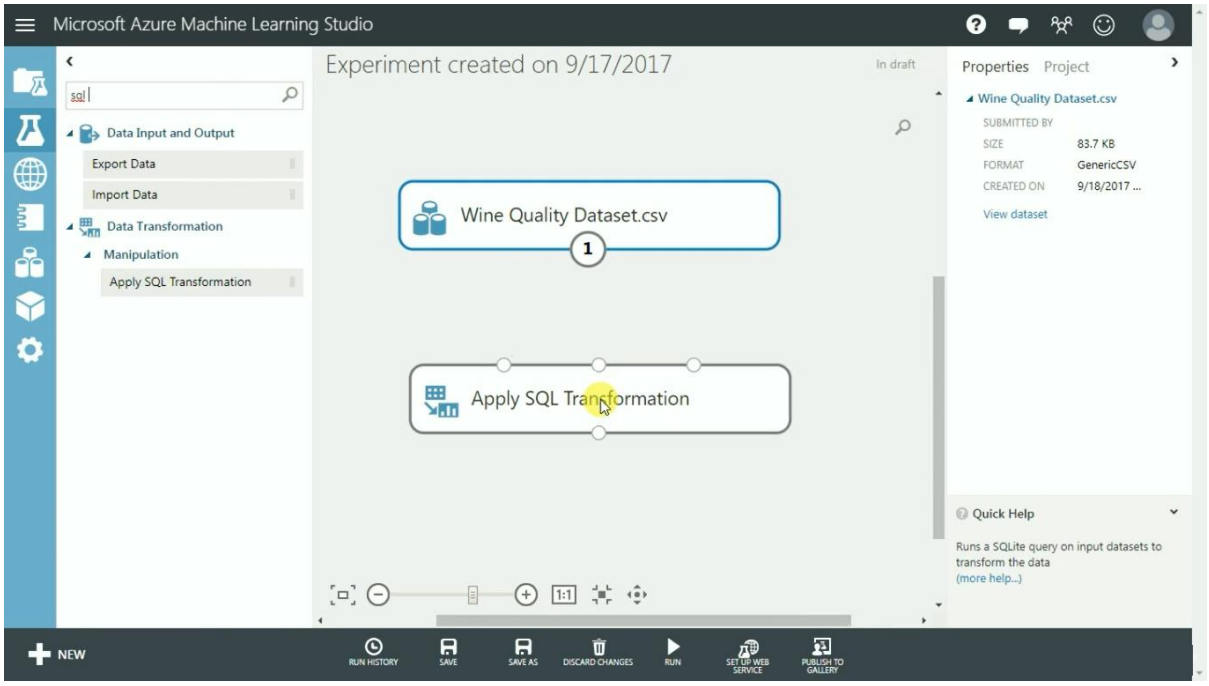
Considering 3 and 4 as low, 5 and 6 as medium, 7 and 8 as high from graph

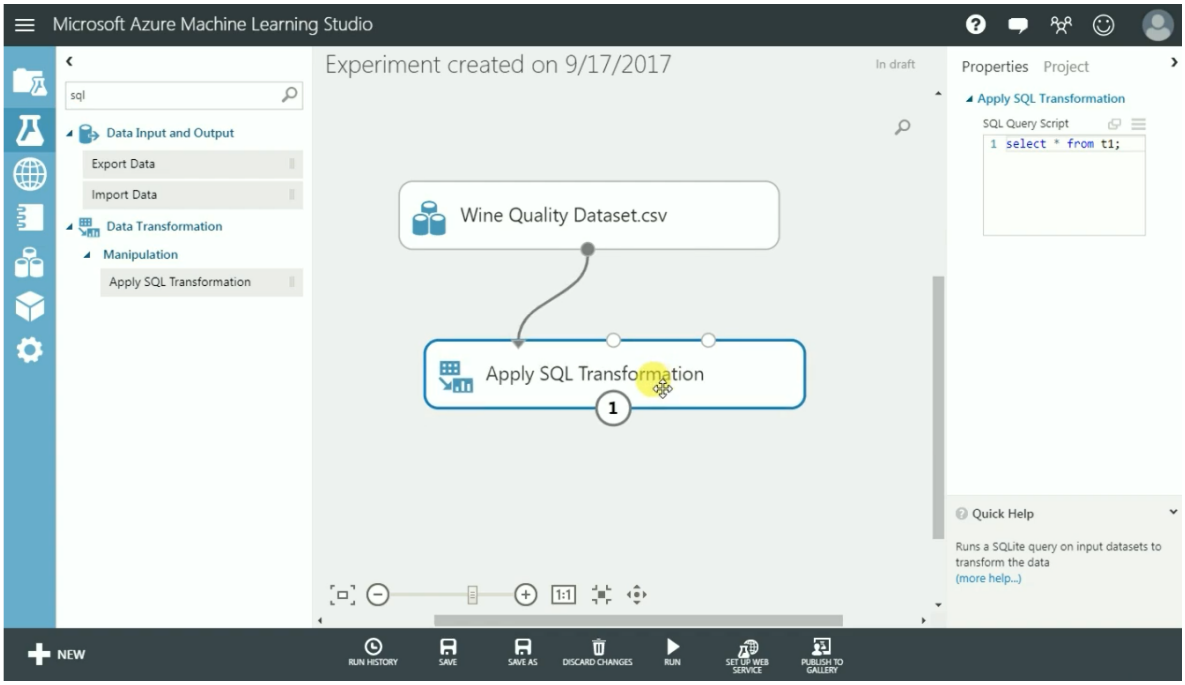


Search for Apply SQL transformation



Connect the nodes



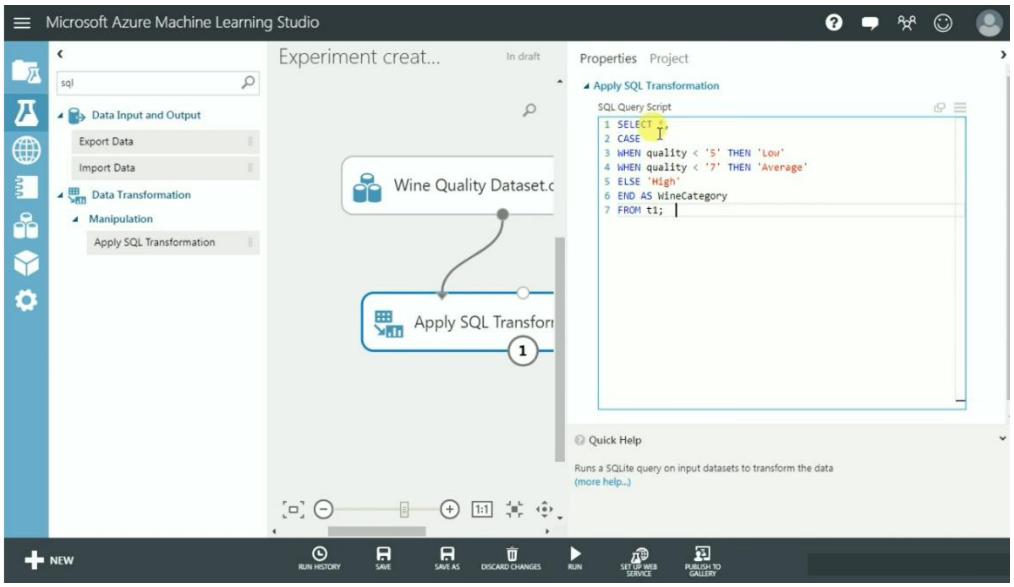


Enter the query script in the right to categorize

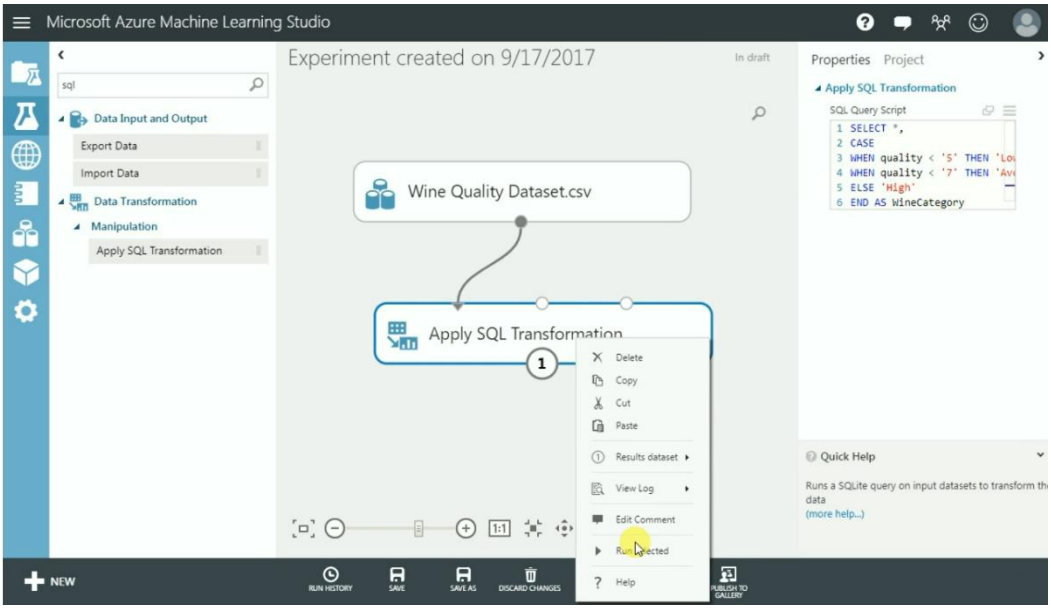
```
SELECT *,
CASE
WHEN quality < '5' THEN 'Low'
WHEN quality < '7' THEN 'Average'
```

ELSE 'High' END AS WineCategory

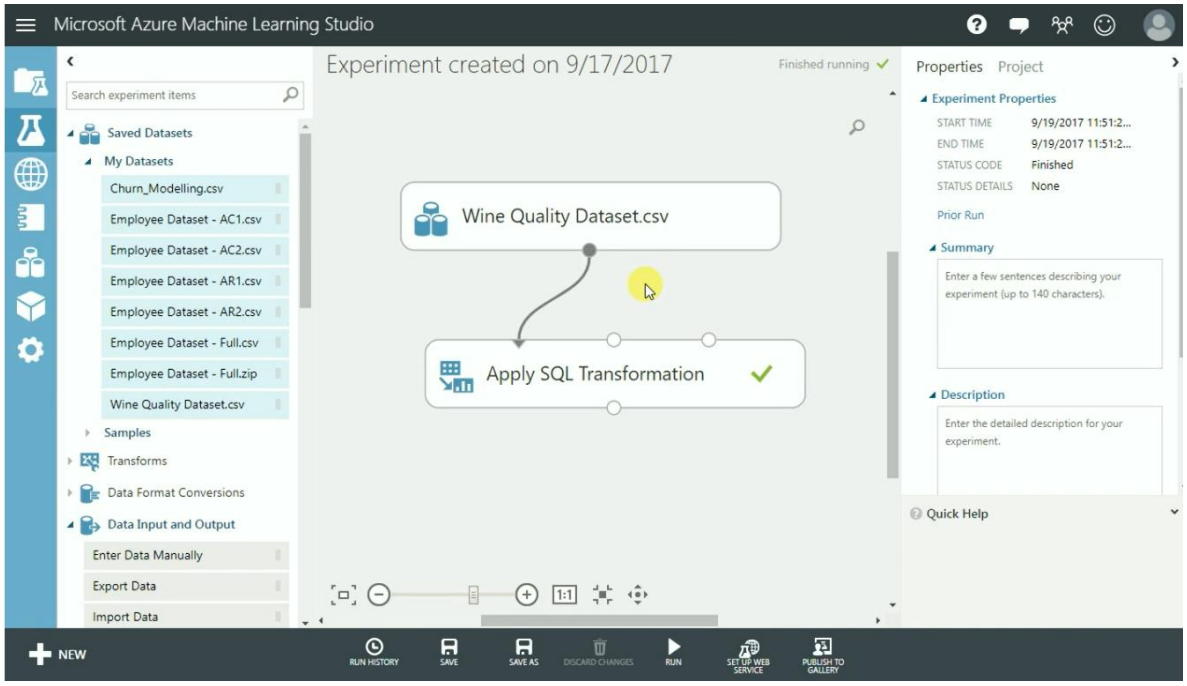
FROM t1;



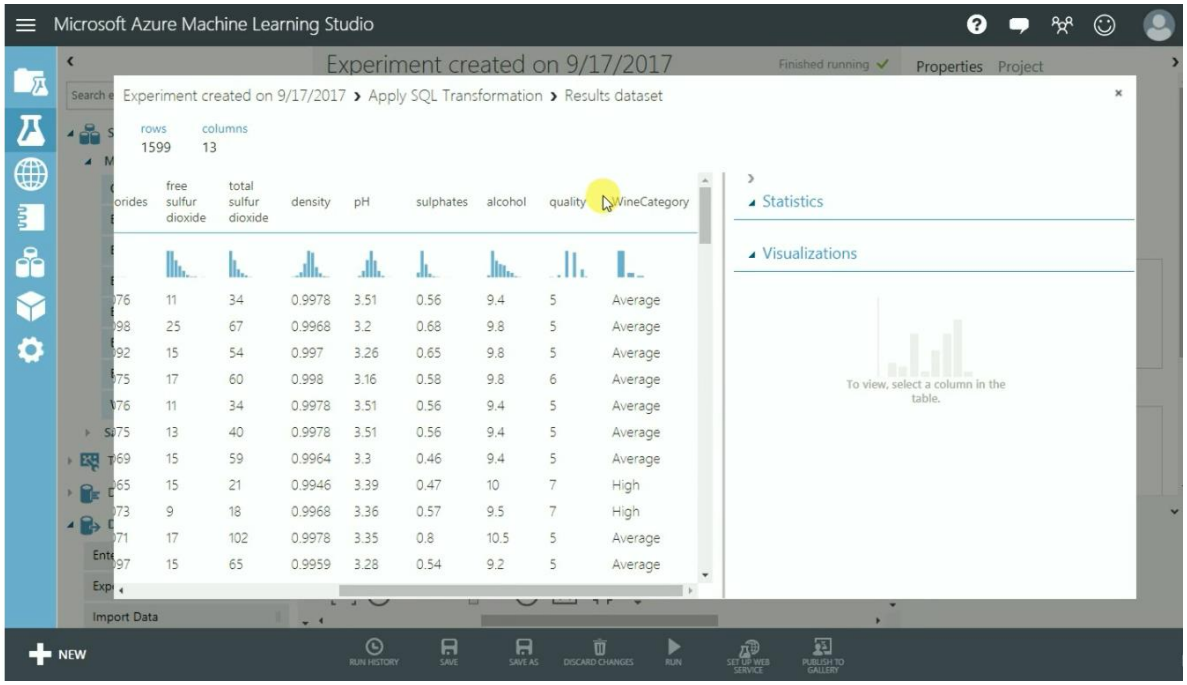
Run for output



Right click and visualize for result

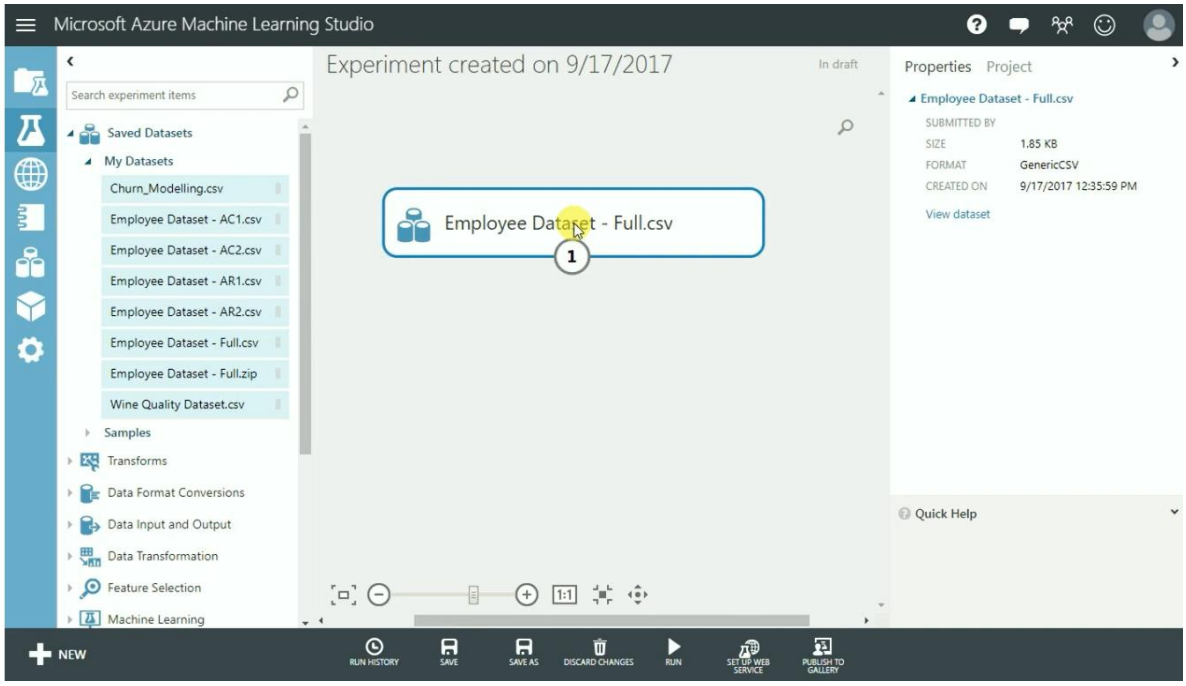


Hence obtained the result successfully using SQL transformation

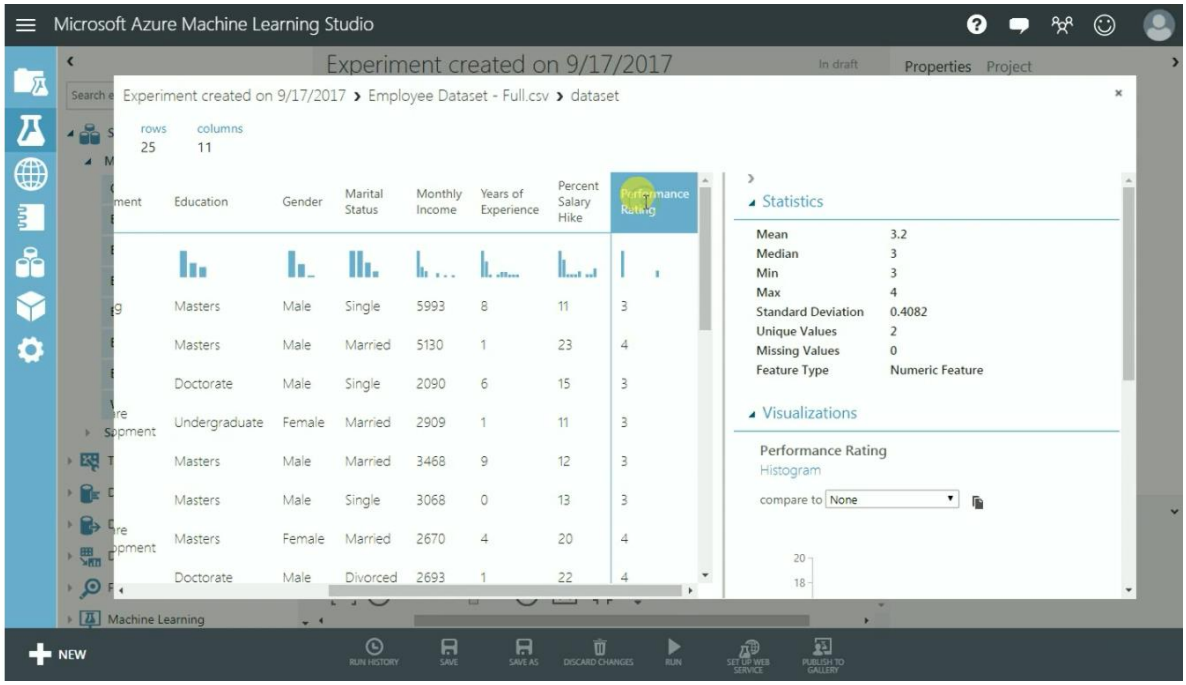


Data Manipulation Using Edit Metadata Component

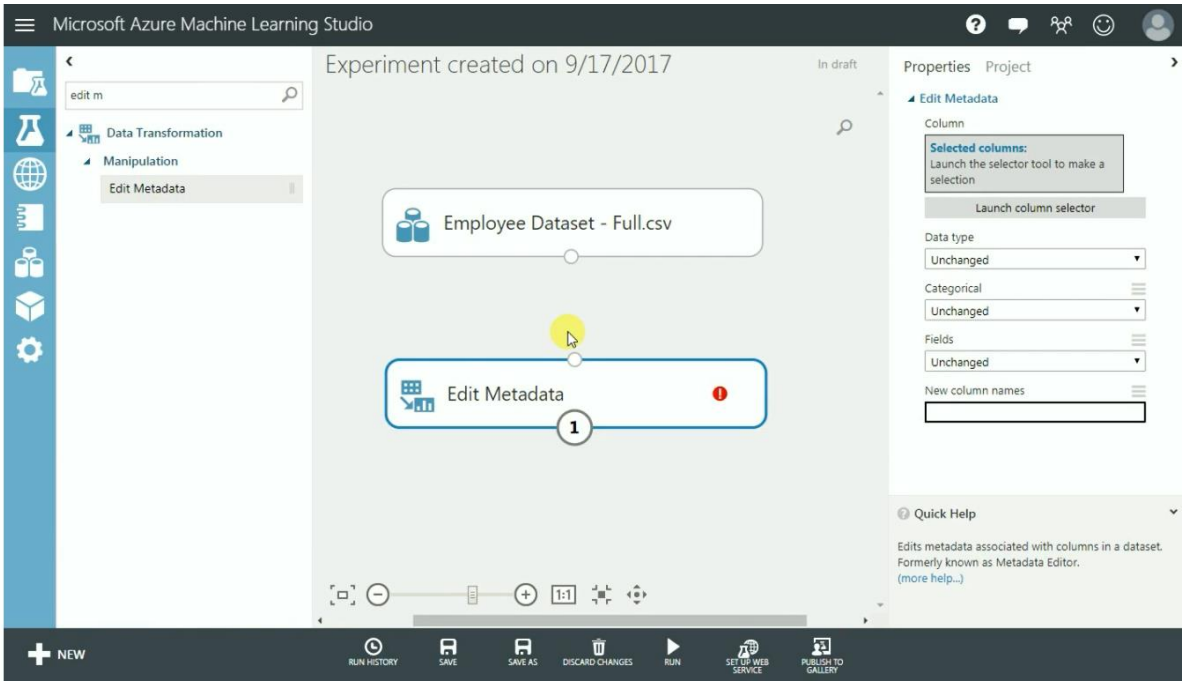
Select and drop the uploaded dataset and visualize



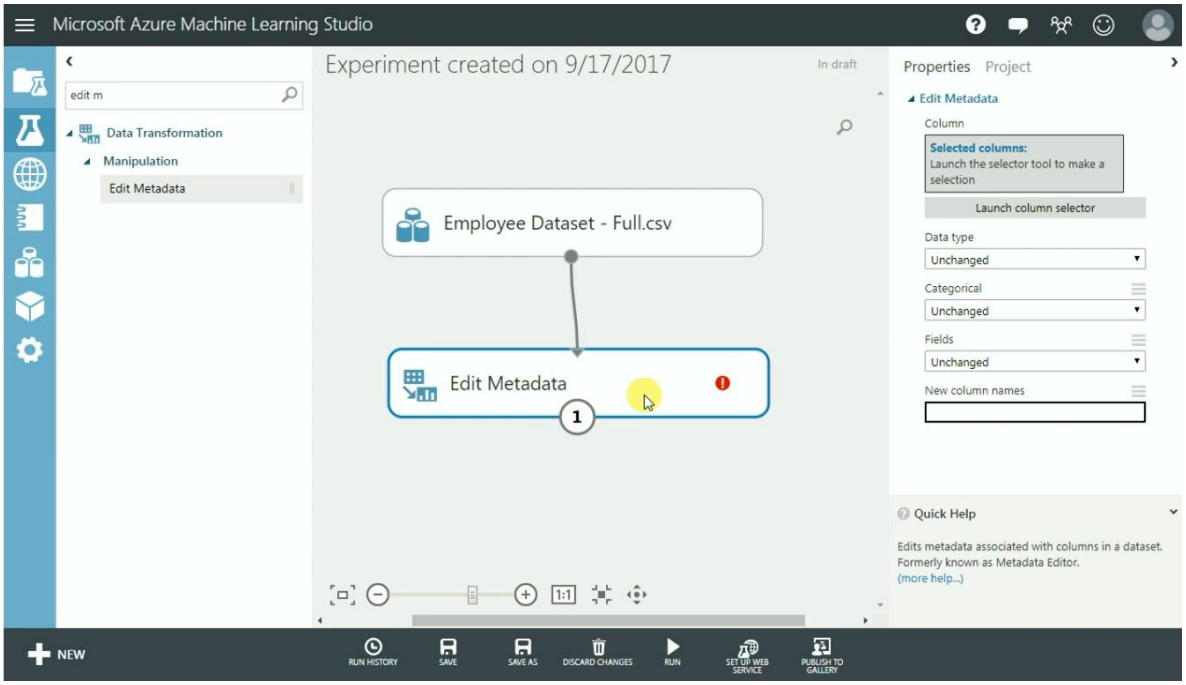
Using metadata can change numeric to category as per requirement



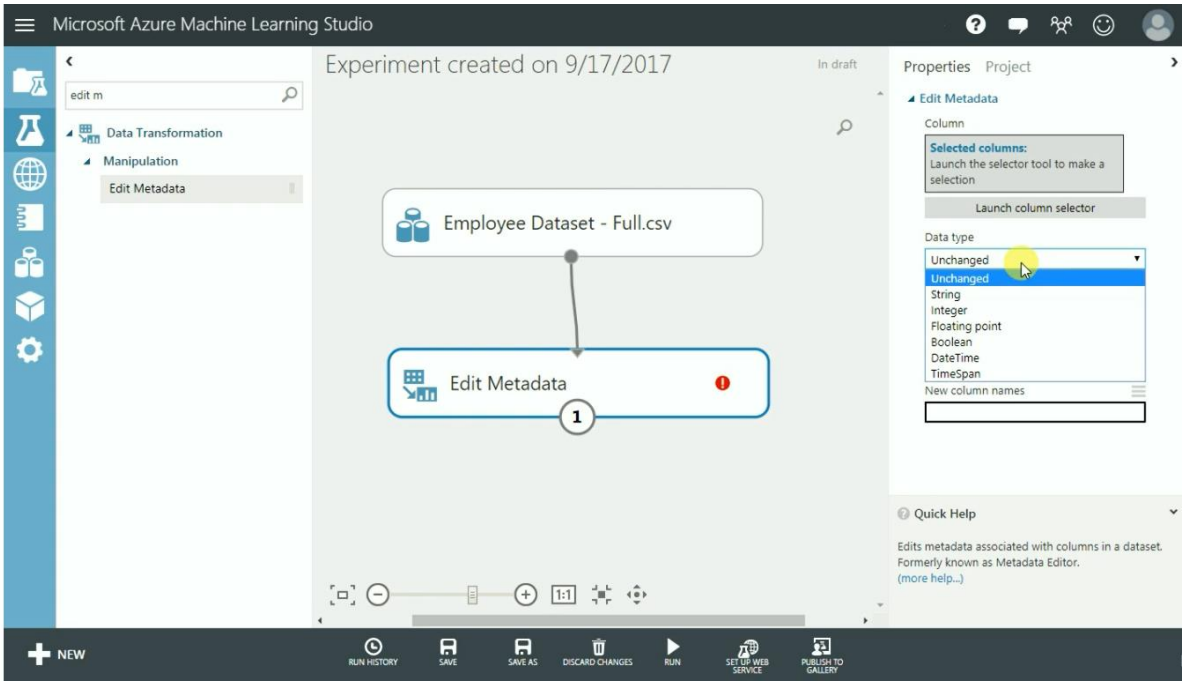
Search for Edit metadata and drop it in the canvas



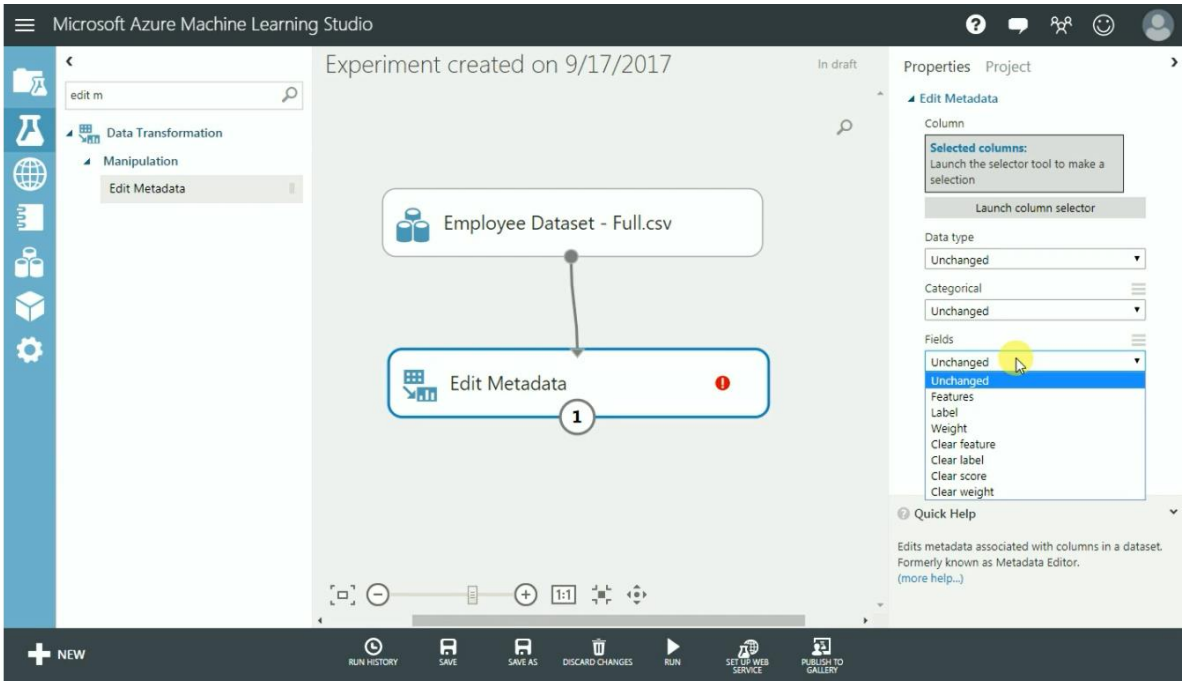
Attach both the data sets



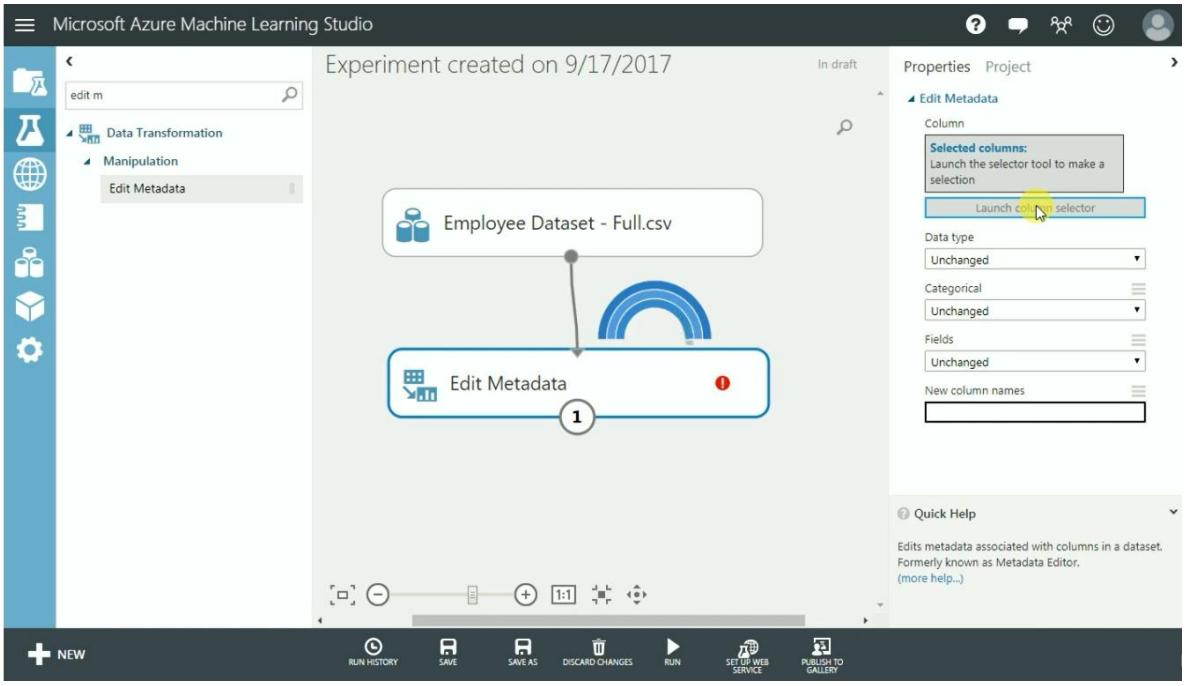
Can change supported data types as per requirement



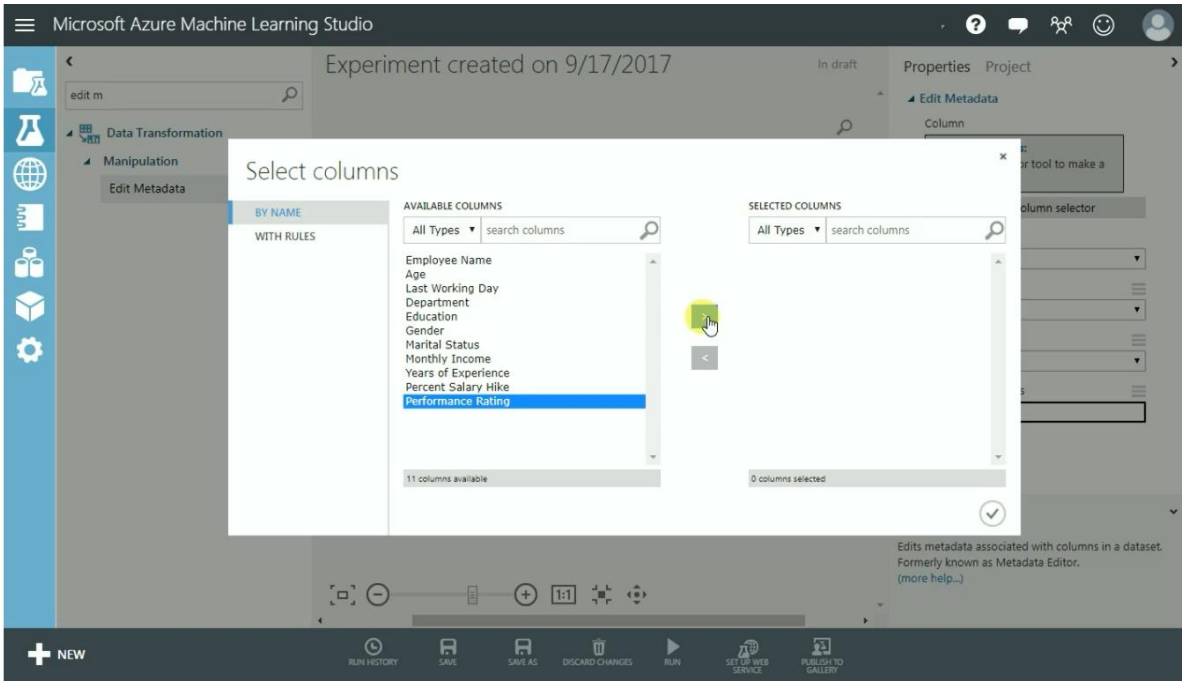
Also can change category and fields in the dropdown



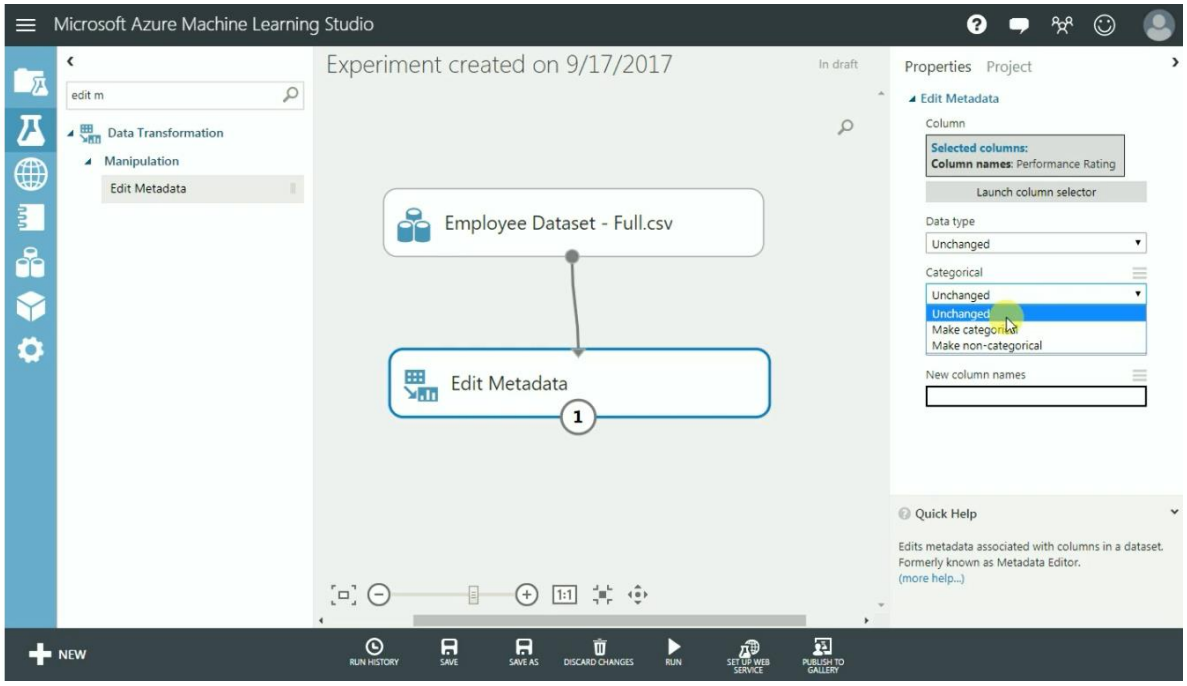
Click launch column sector to select the column in which metadata to be changed



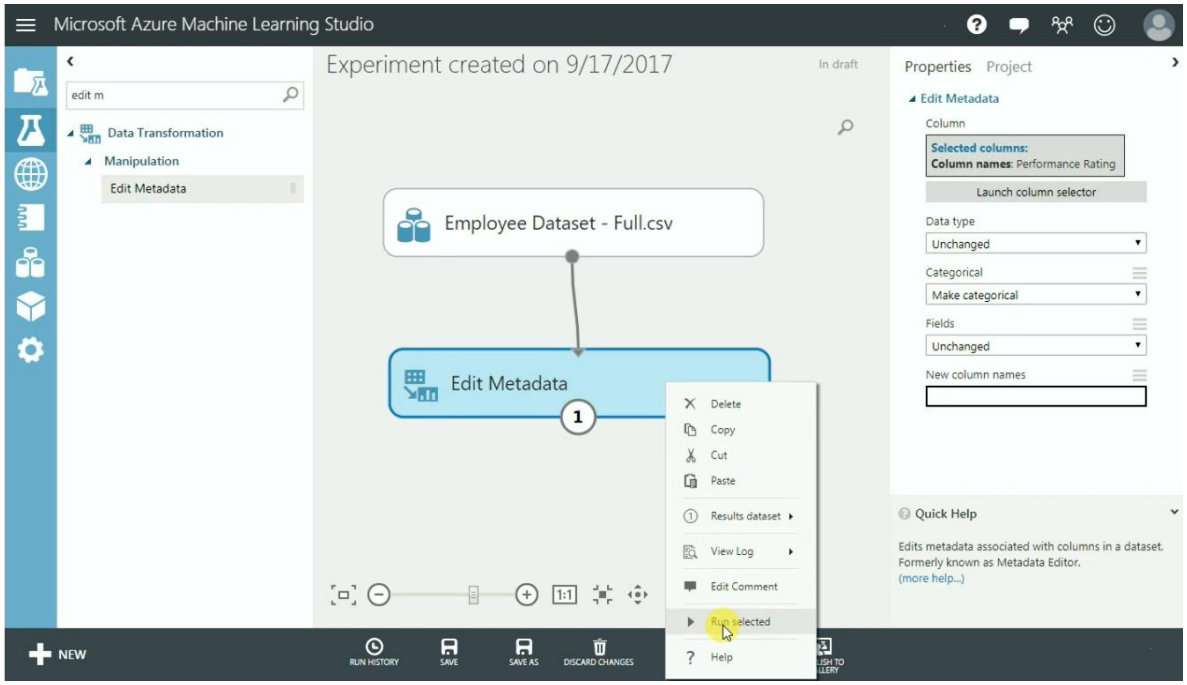
In this case select performance rating to change the metadata and click ok



Select make categorical from dropdown

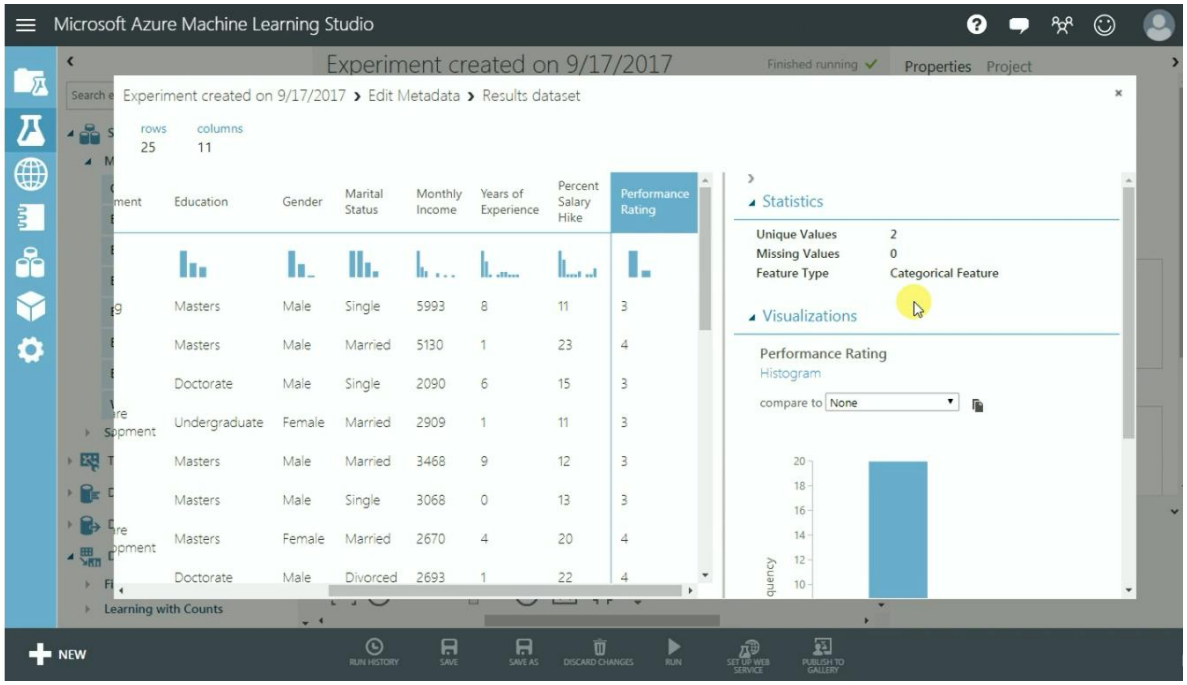


Run for obtaining result and click visualize after completion



Go to our column performance rating and select

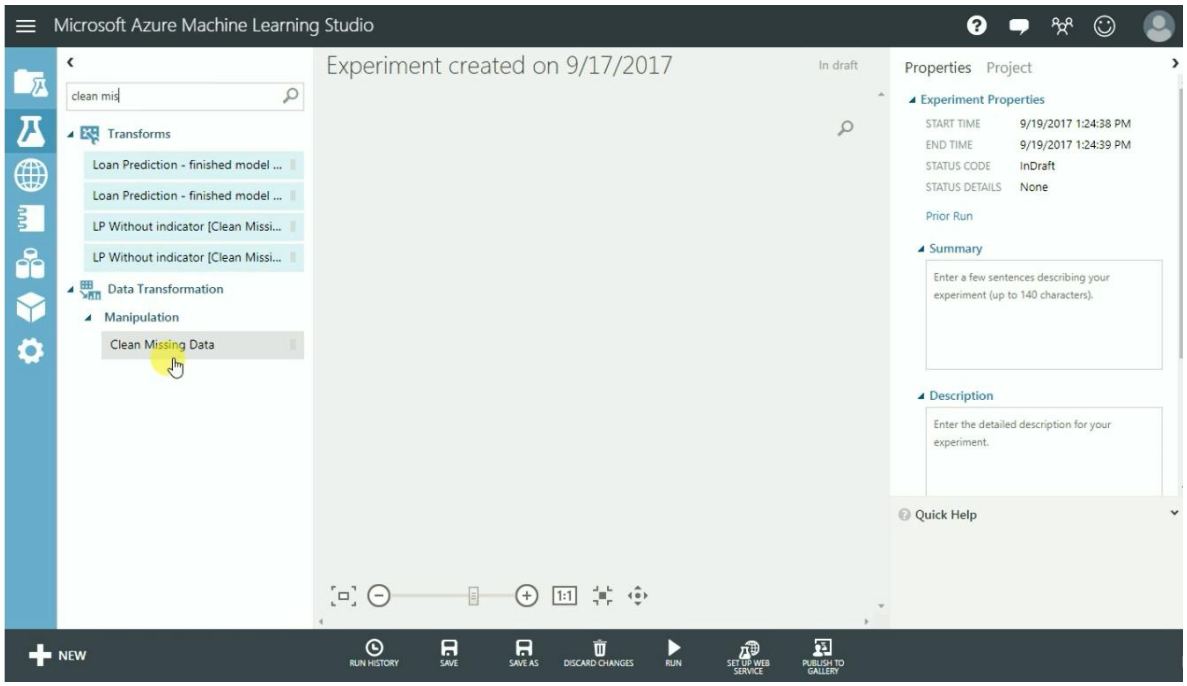
Now you can see the categorical feature type successfully



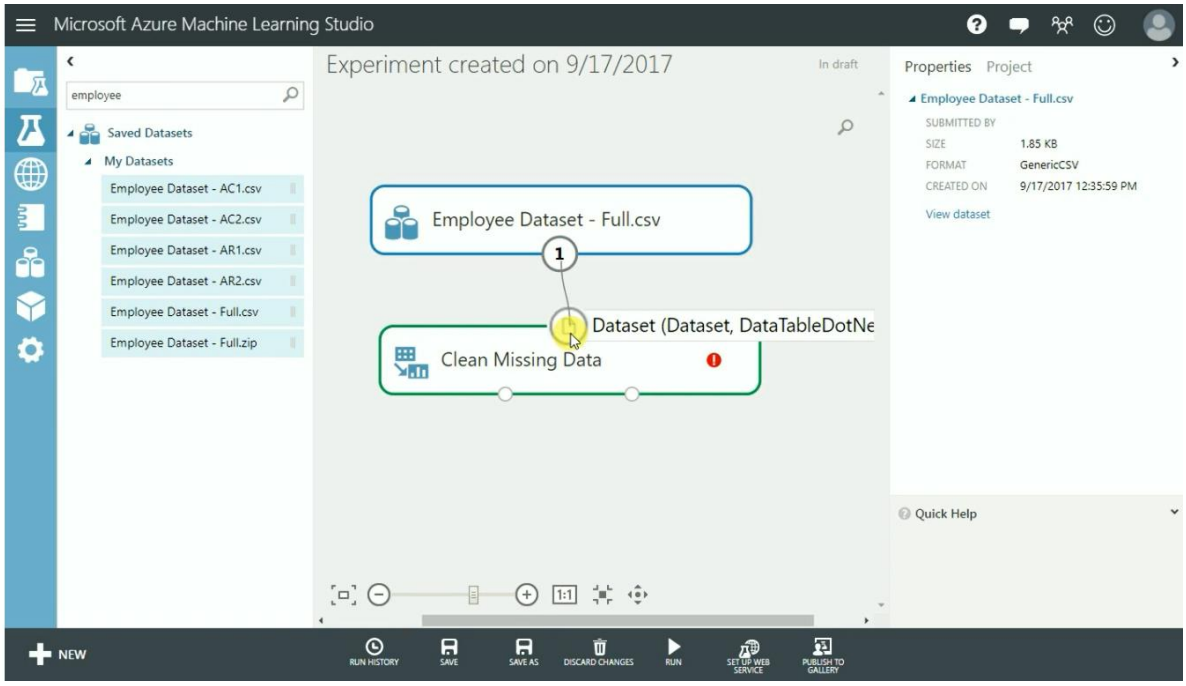
Data Manipulation Using Clean Missing Data Component

CLEAN MISSING DATA

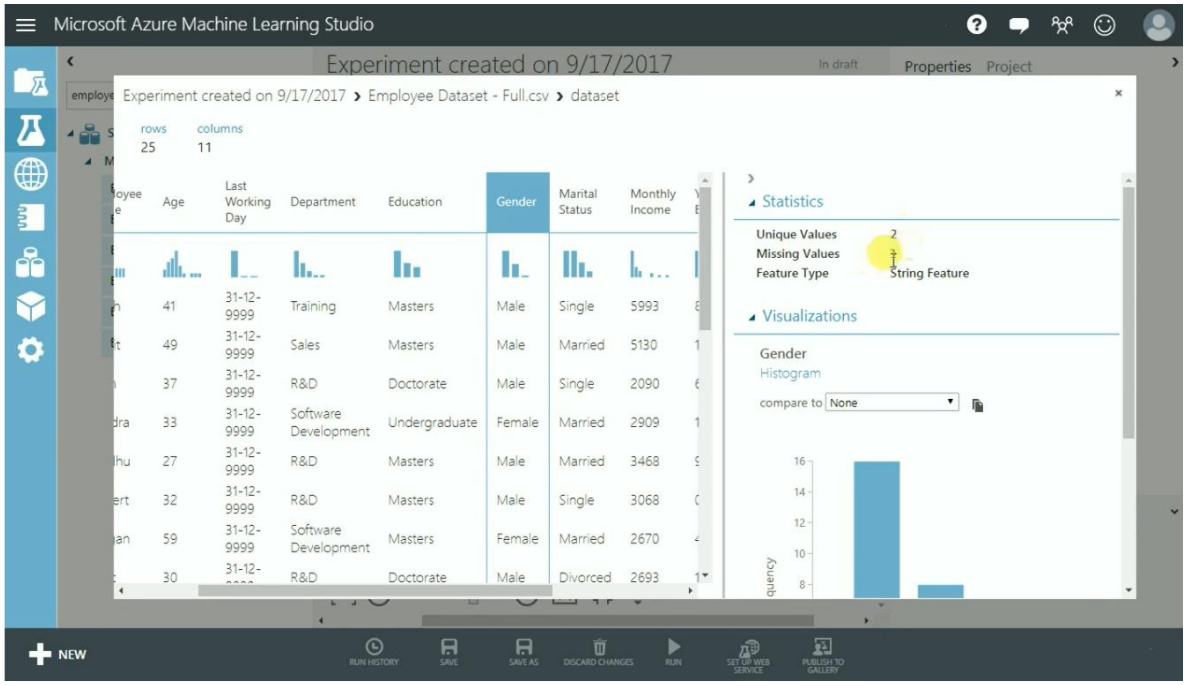
Search for clean missing data and drop in the canvas



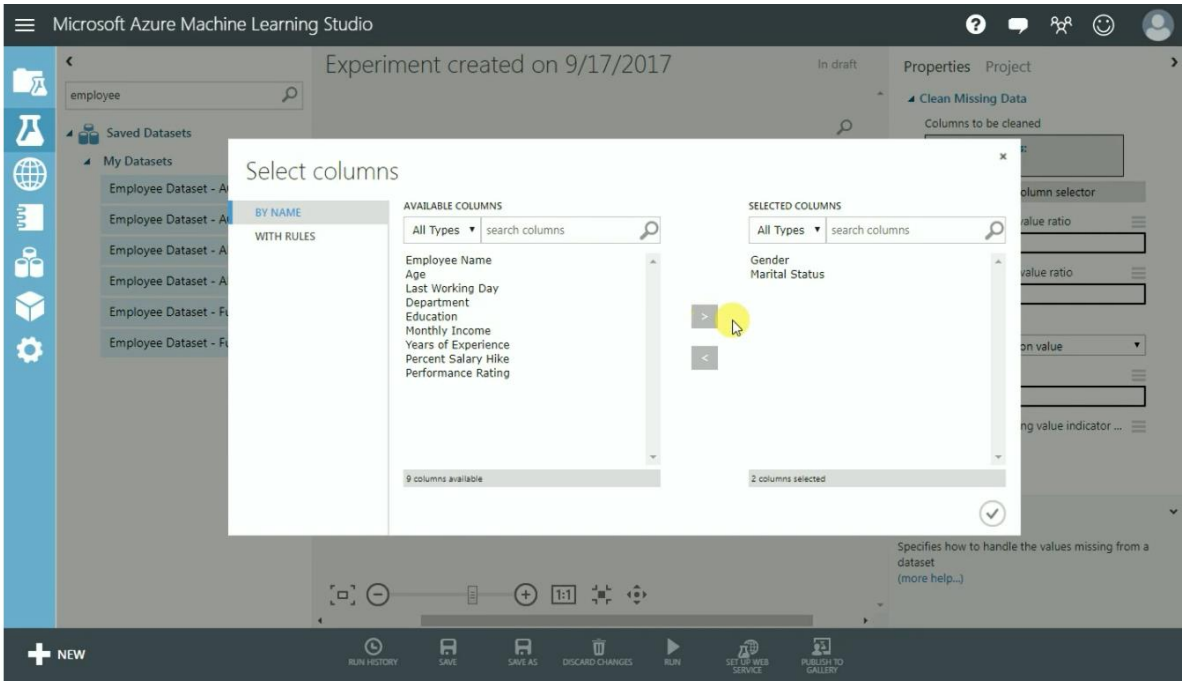
Also take the inputted dataset into canvas and attach with clean missing data



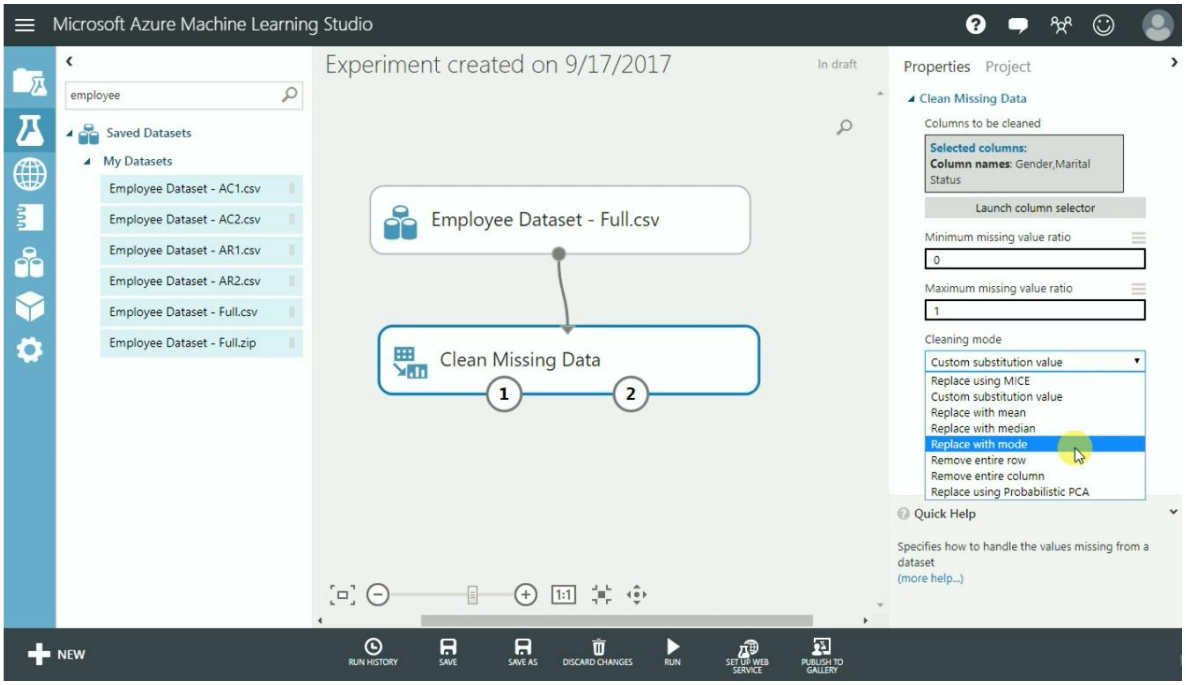
Visualize the dataset and check for any missing values in each columns



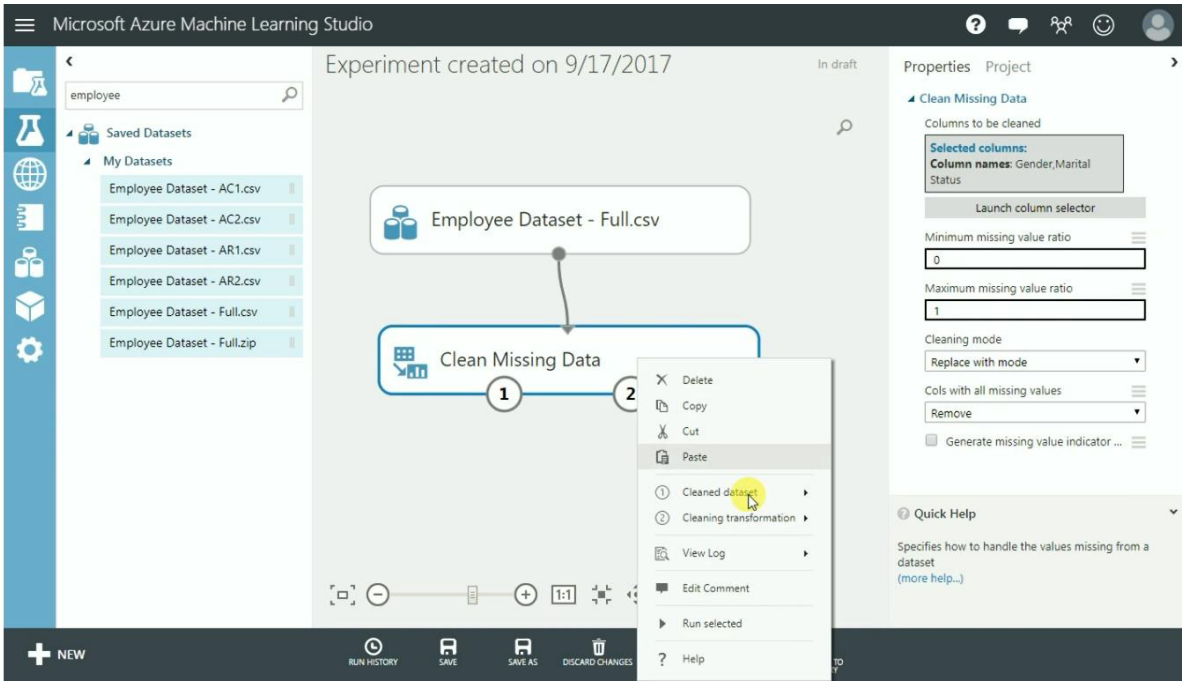
Launch column selector and select the two columns which has missing values



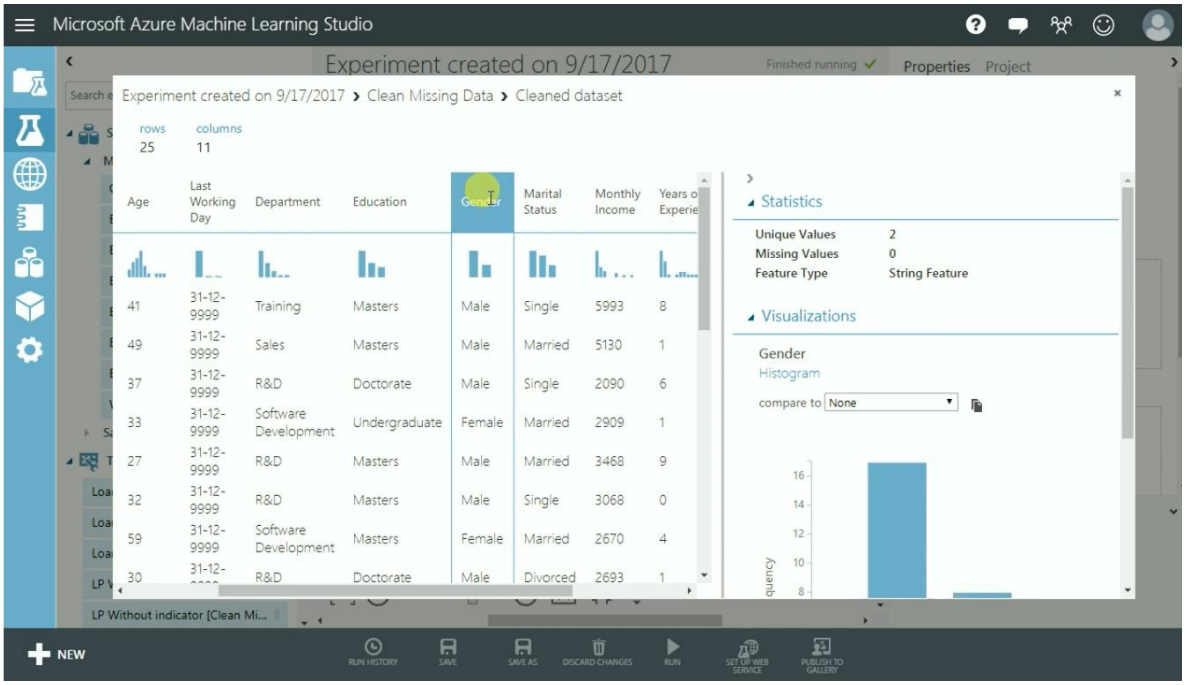
Select Replace with mode from dropdown



Run the module

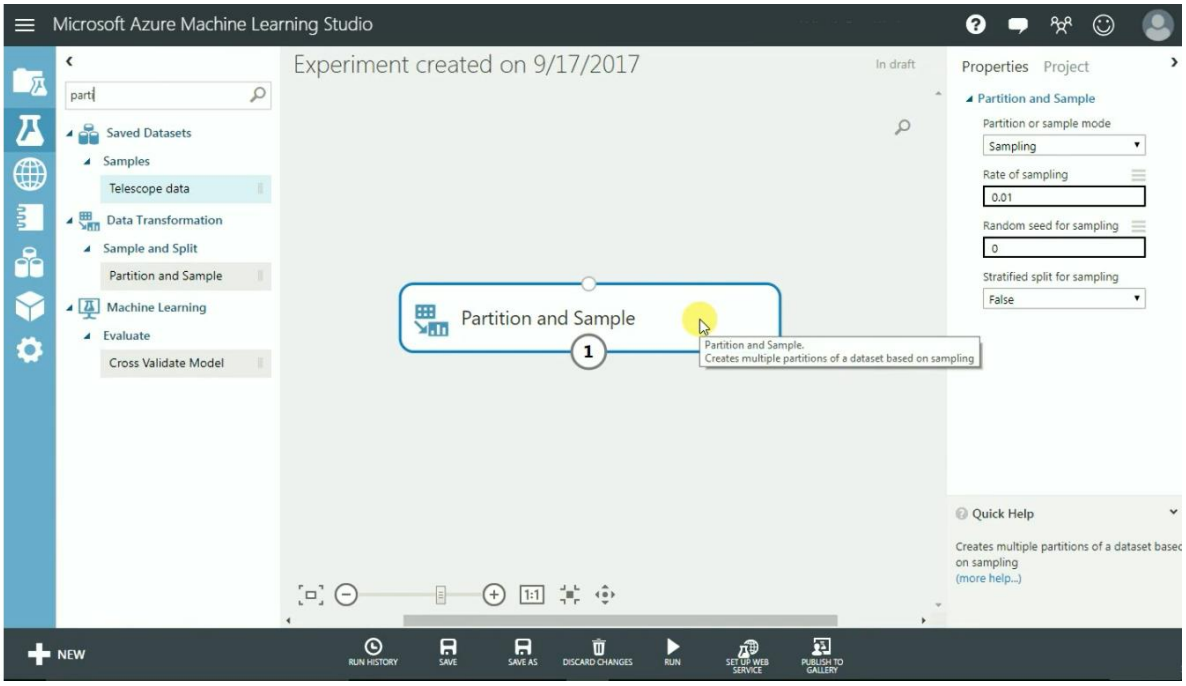


After execution visualize to check the result, Now you can see that there are no missing values in the selected columns and shows clean missing data successful



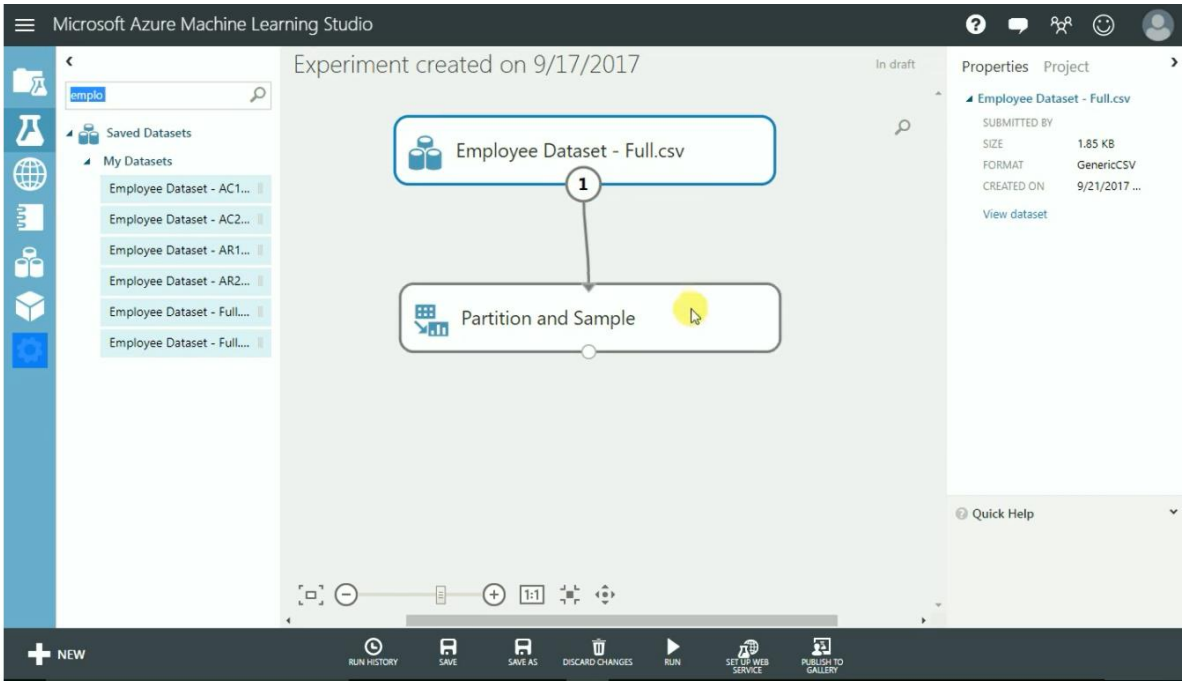
Data Manipulation Using Partition & Sampling Component

Search for partition and sample and drop in canvas



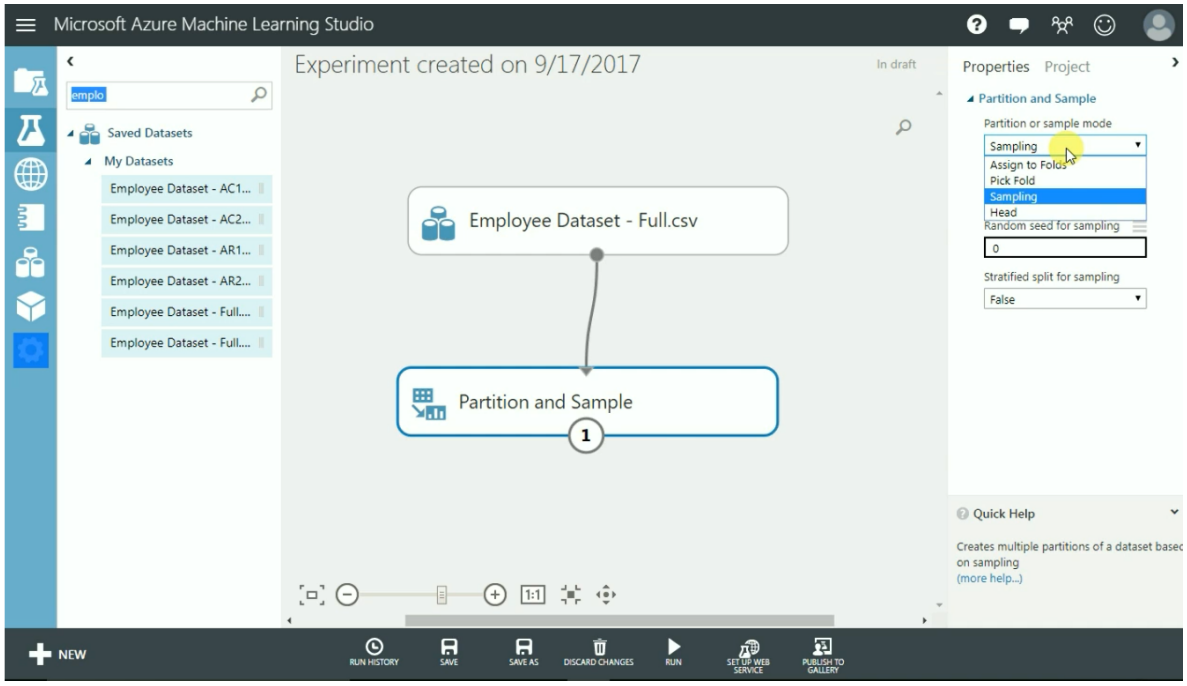
To perform partition and sample operation add an existing dataset in canvas

And connect the nodes as required

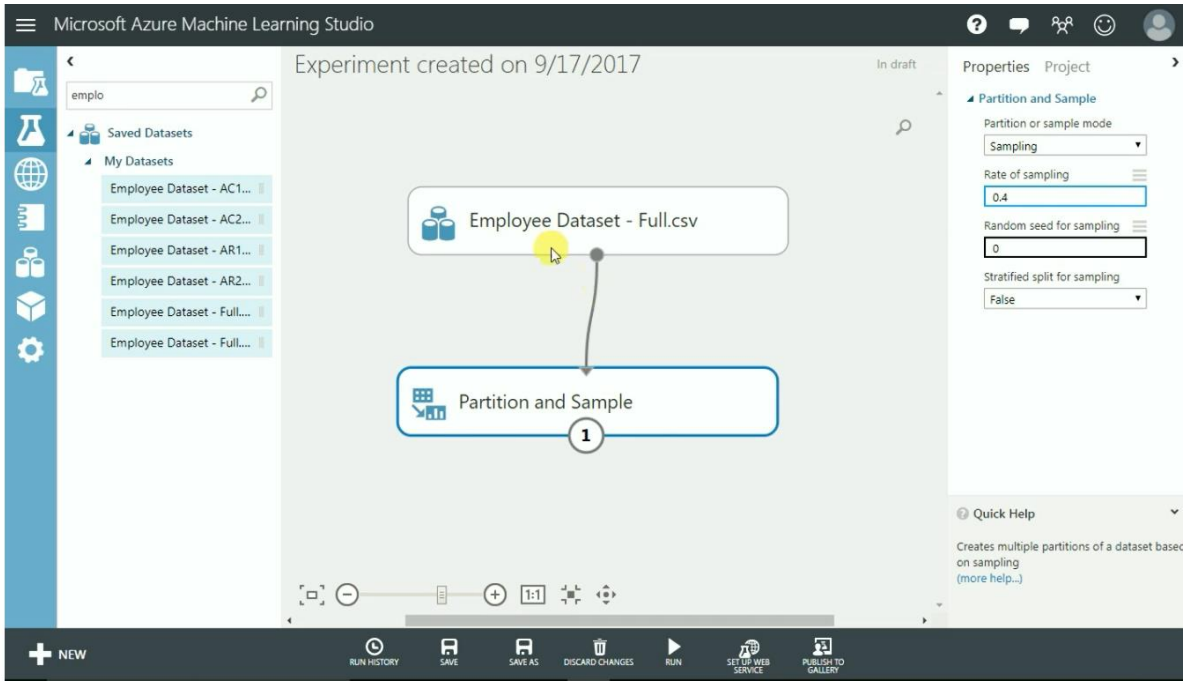


Can apply any of the types from dropdown in which we can apply

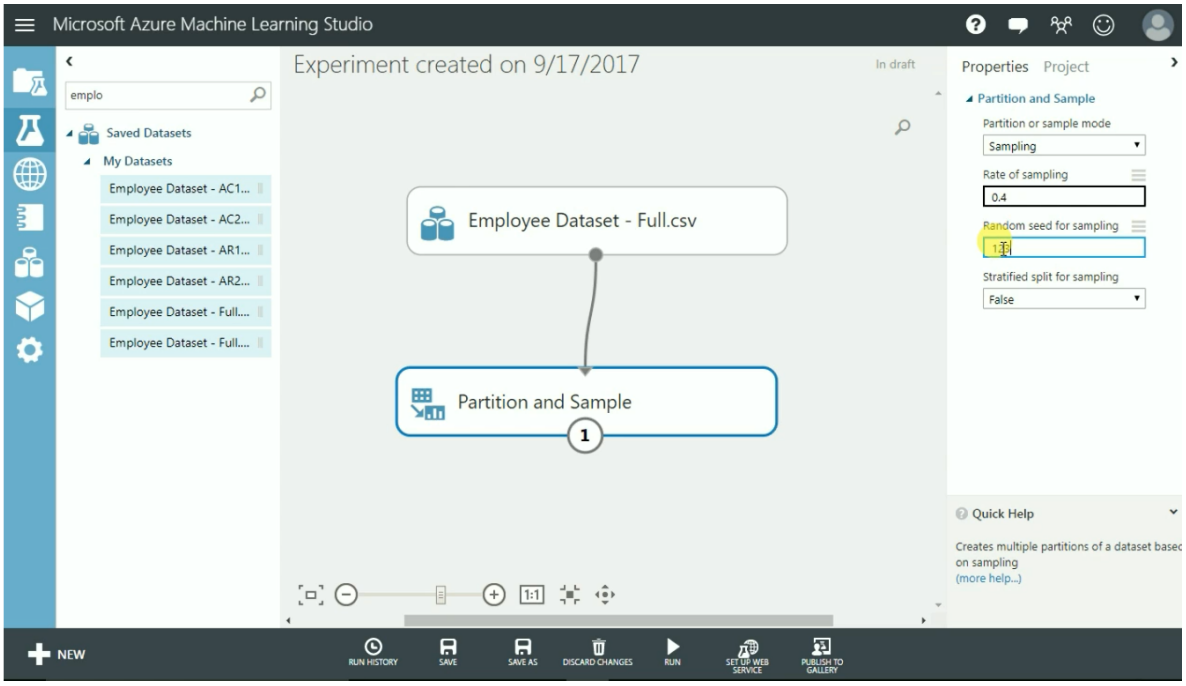
Sampling from the list



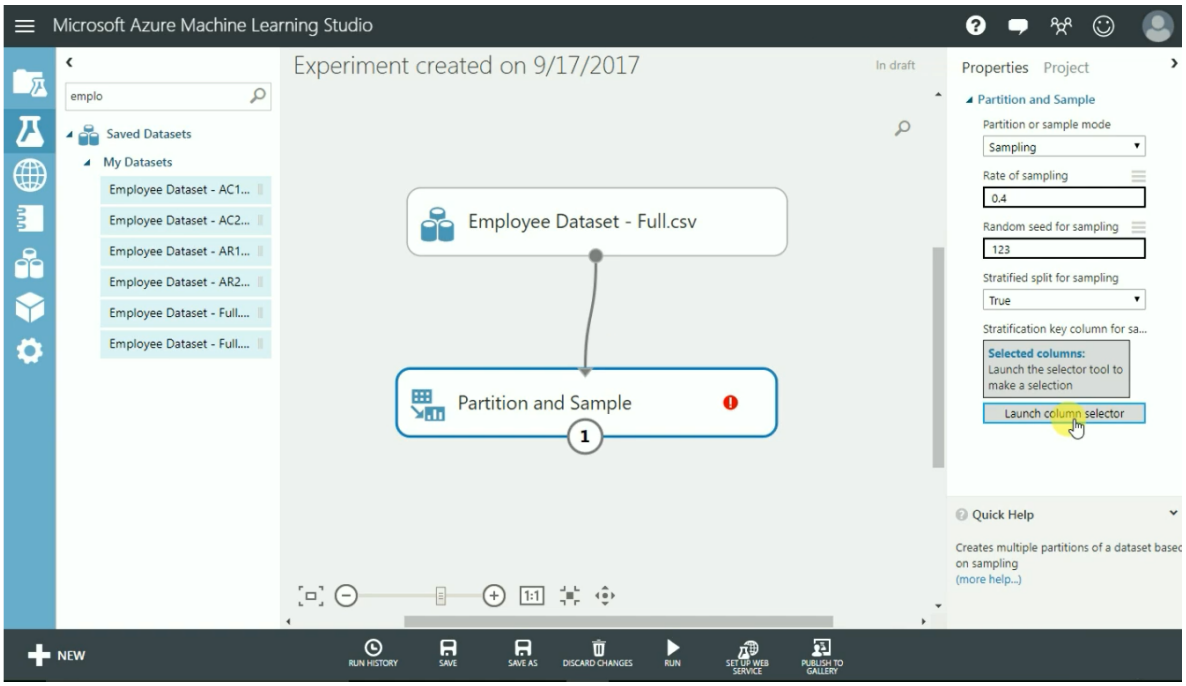
Enter the rate of sampling



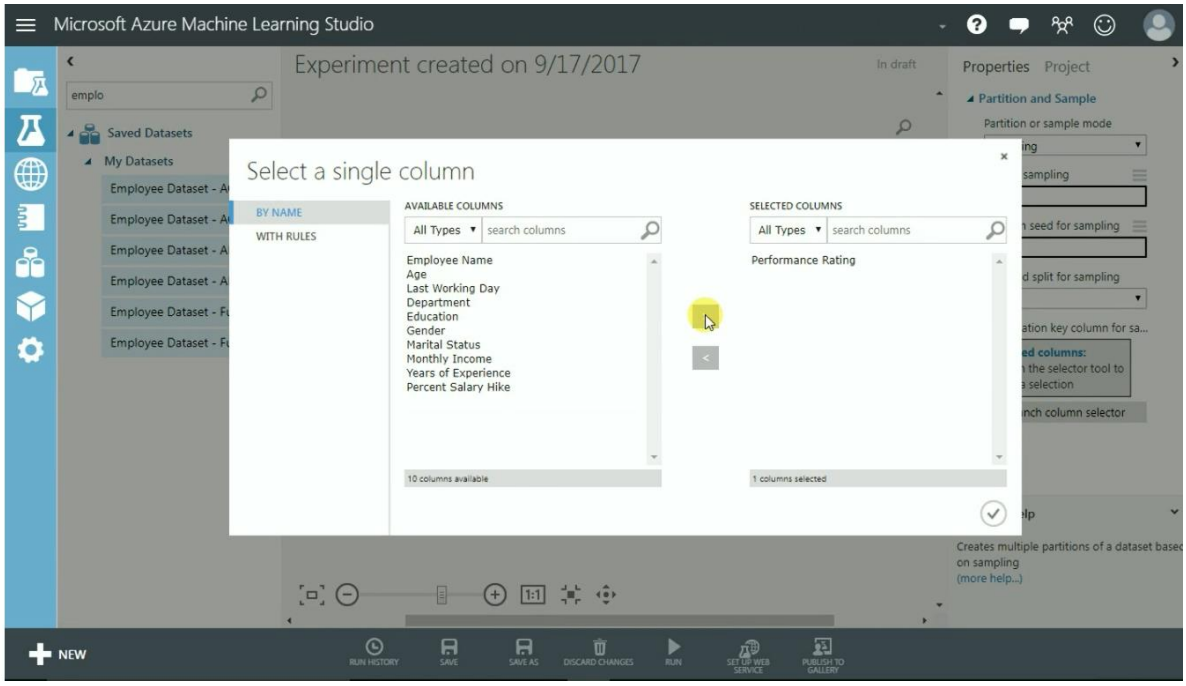
Input random seed for sampling



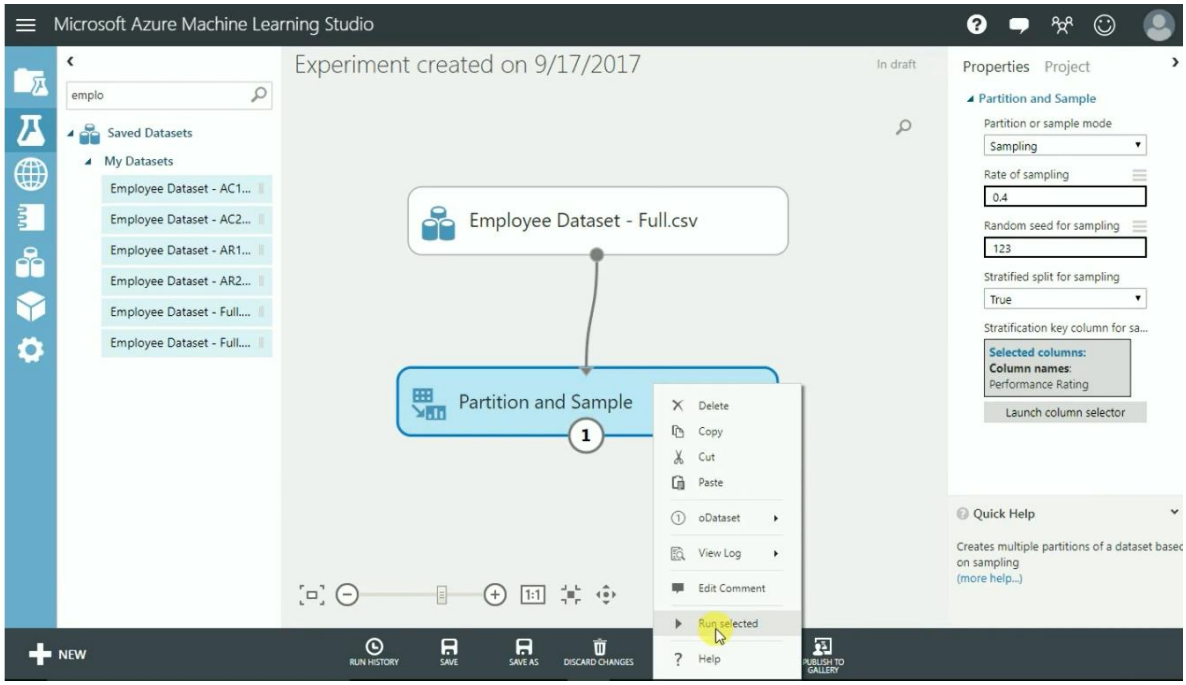
Input stratified split for sampling as true and then select launch columns selector



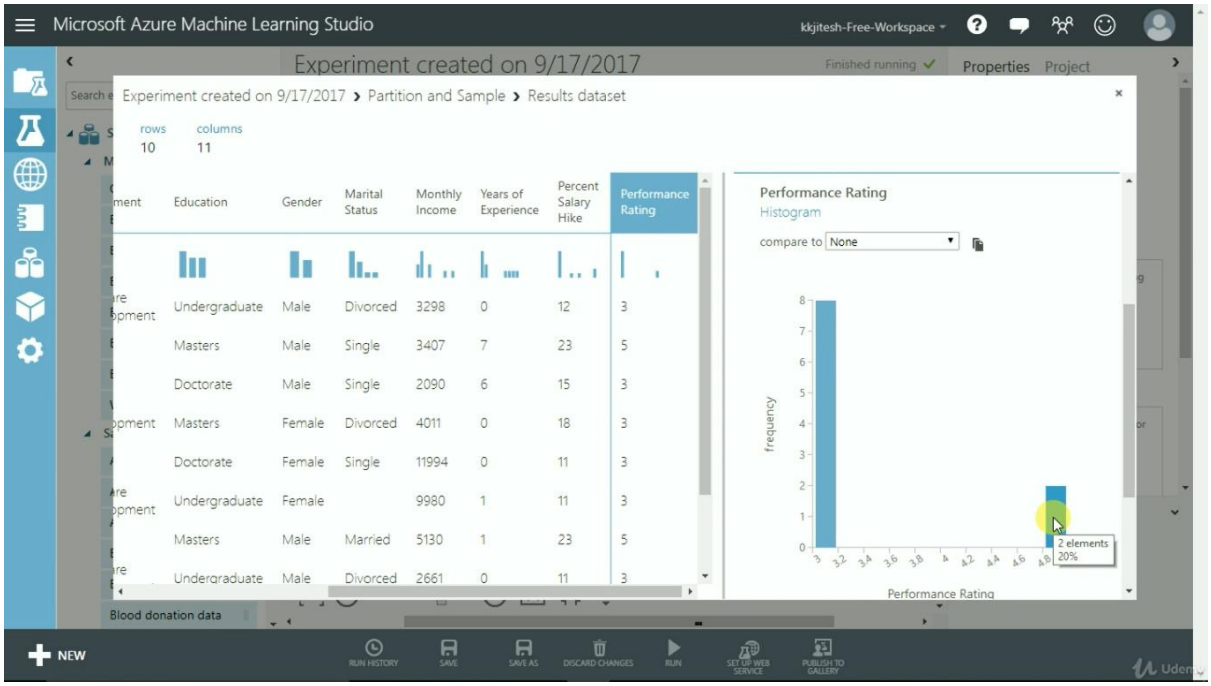
Select performance rating to the selected columns and click ok



Run the module

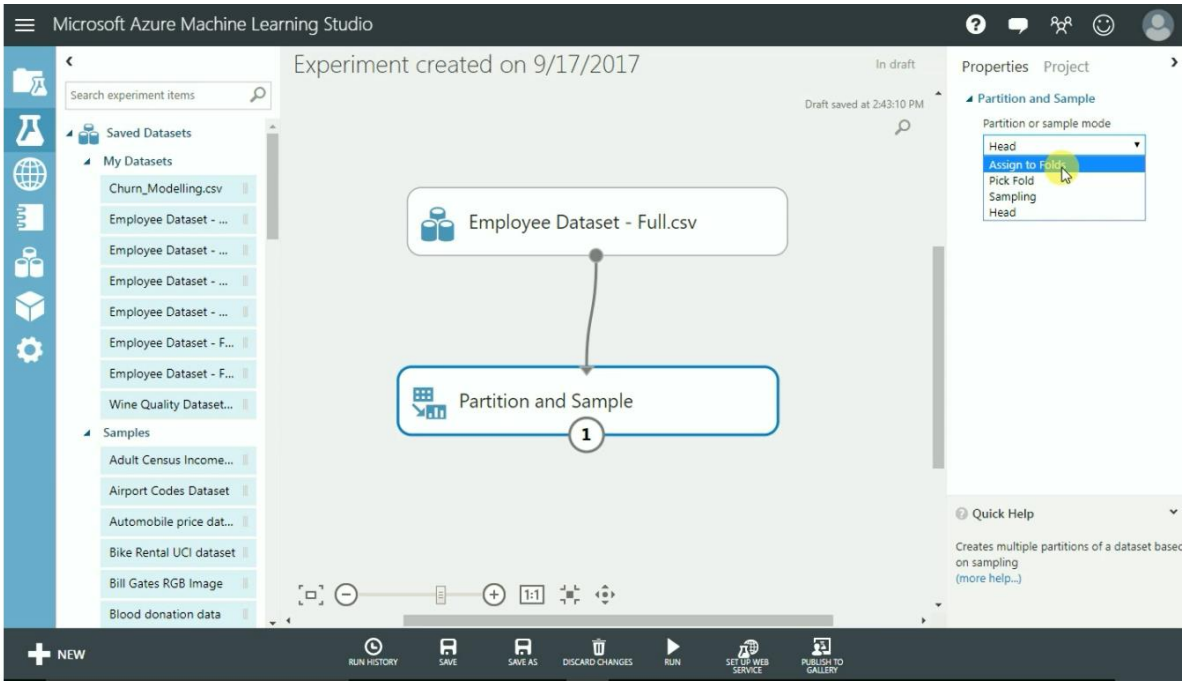


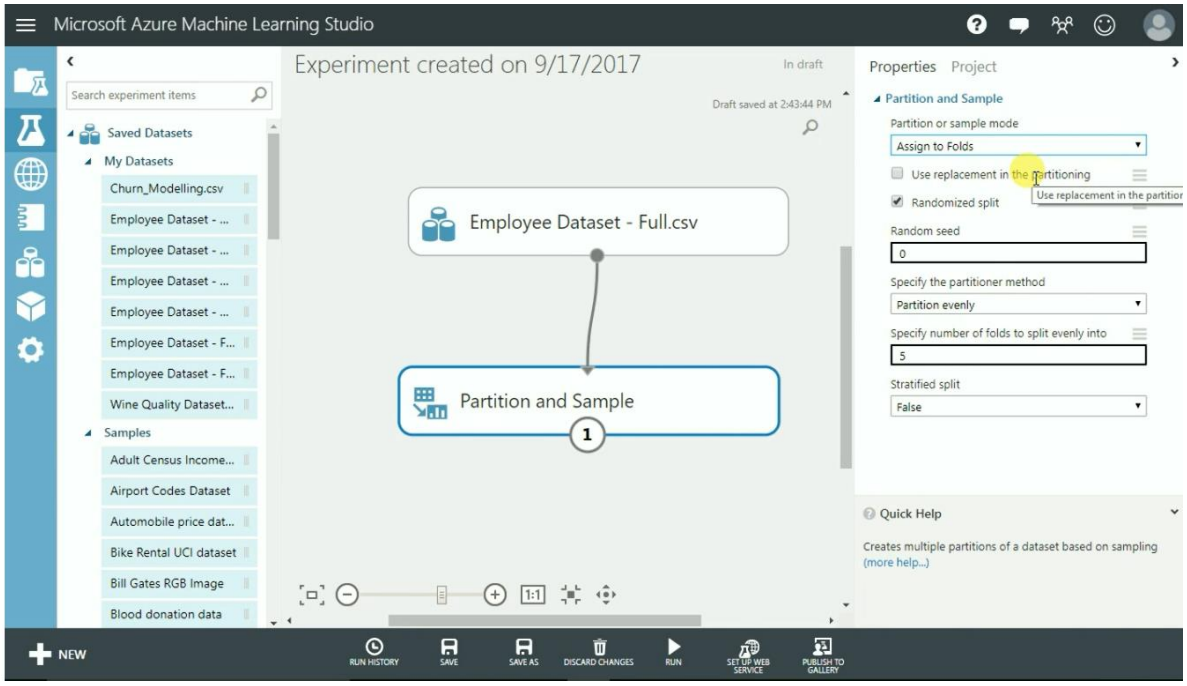
Visualize the successful output



Another option of sample mode

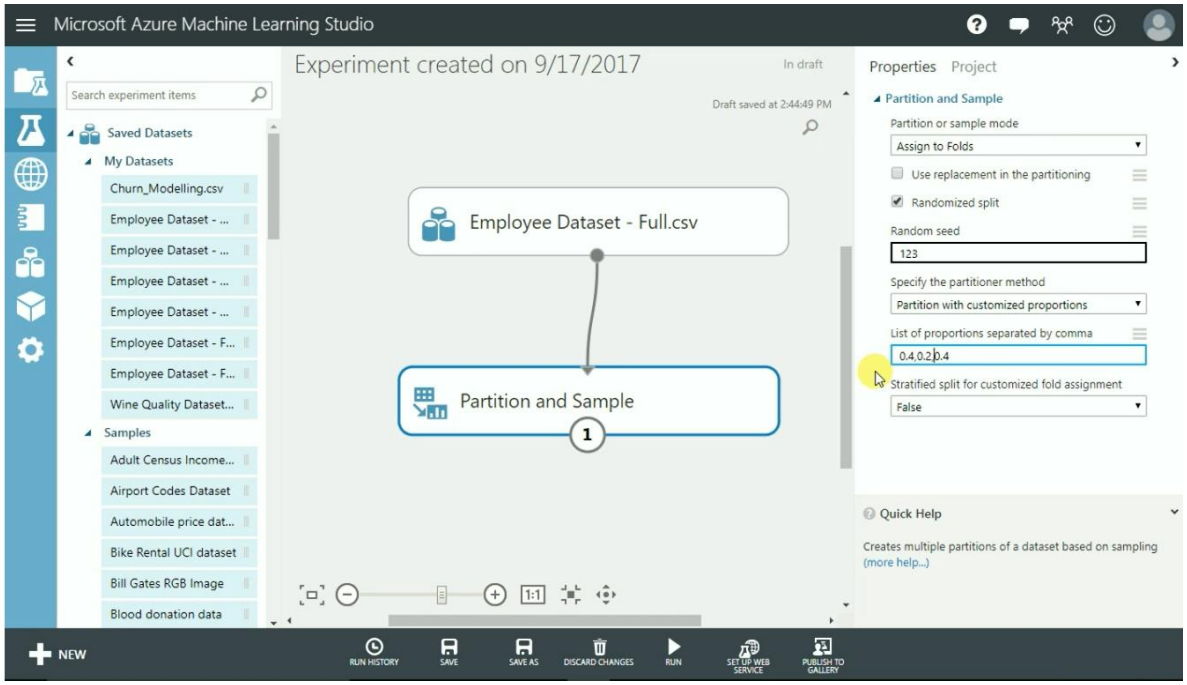
From the existing dataset , select assign to folds from dropdown



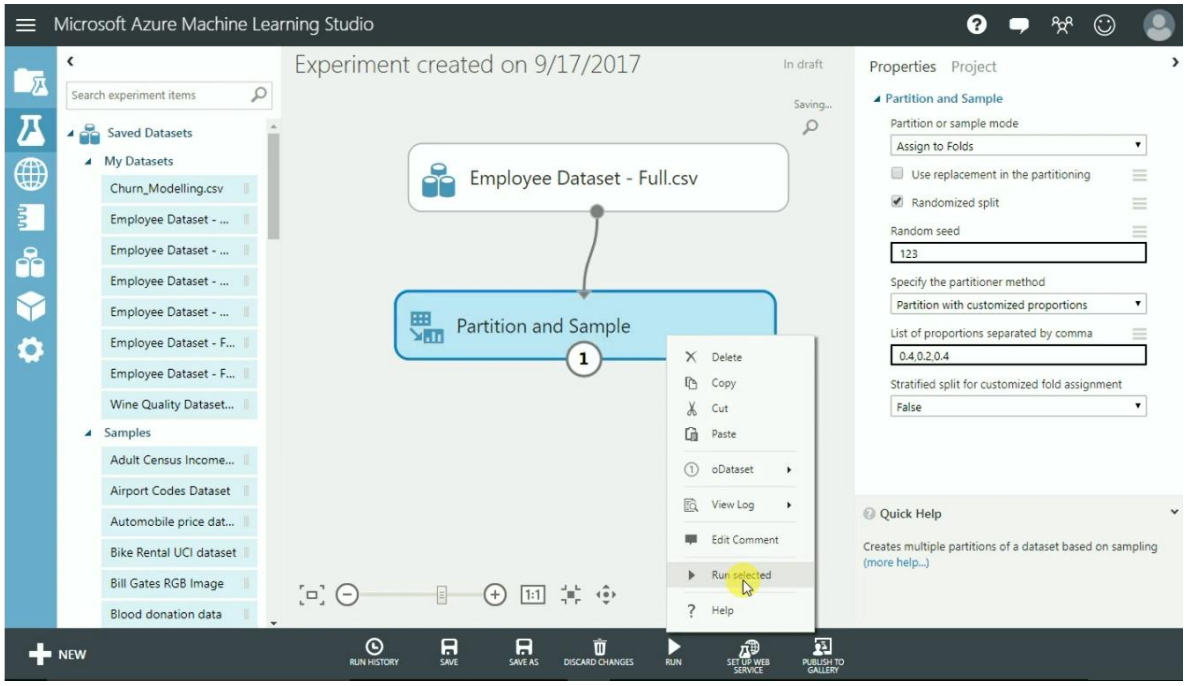


Select partition with customized proportions and list the partitions

Required separated by commas

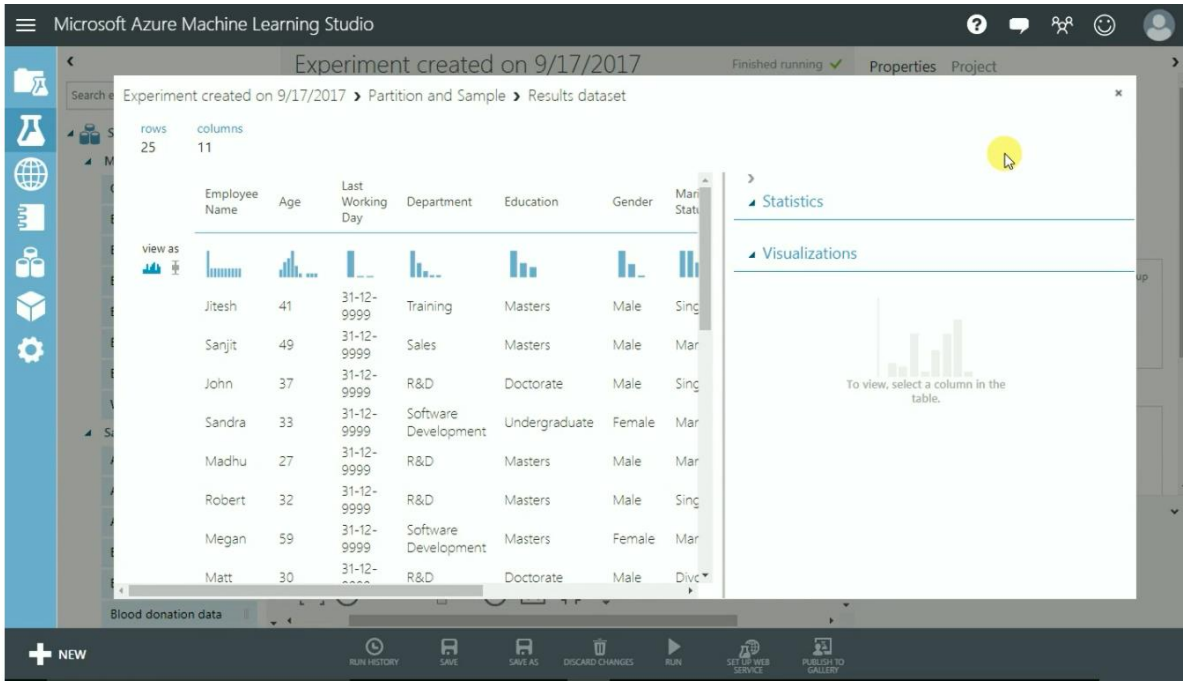


Run the module

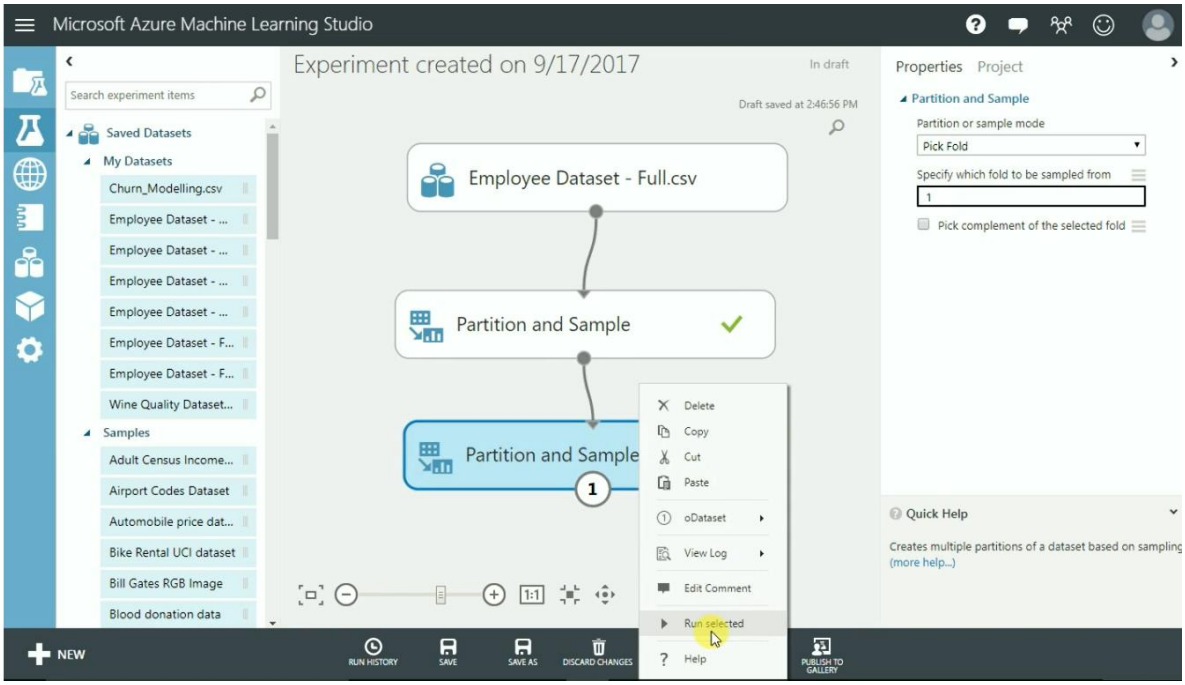


Now the visualization shows a single dataset not as three different dataset

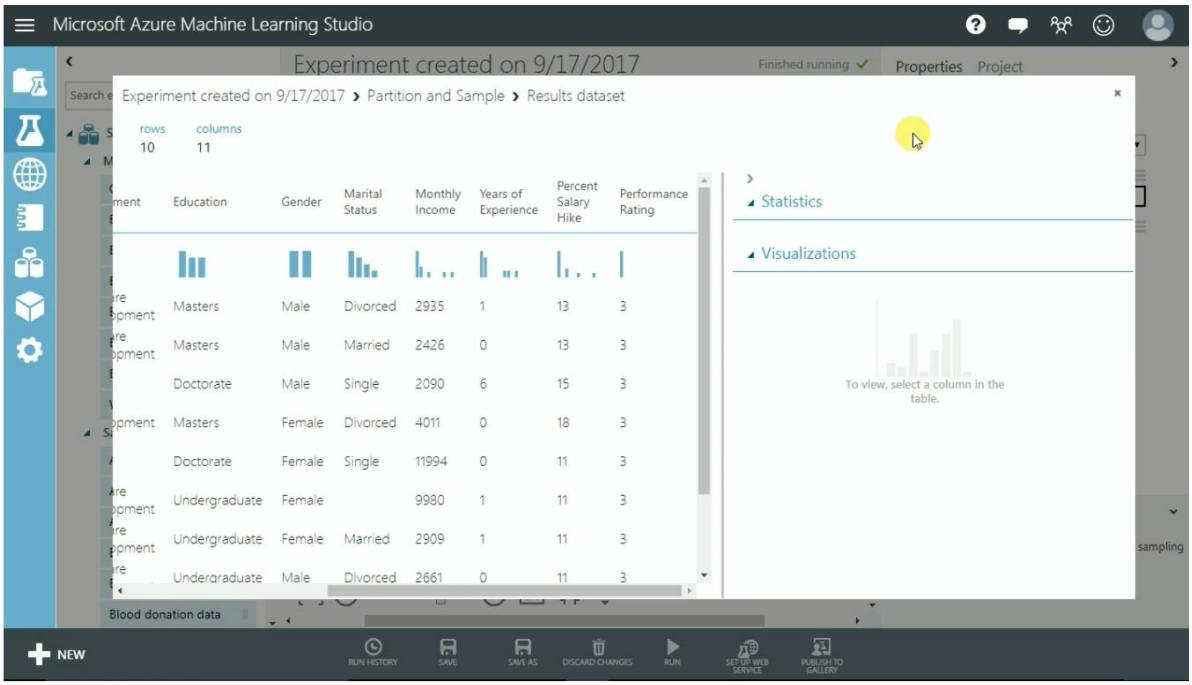
Microsoft should provide other visualization for this module alone



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

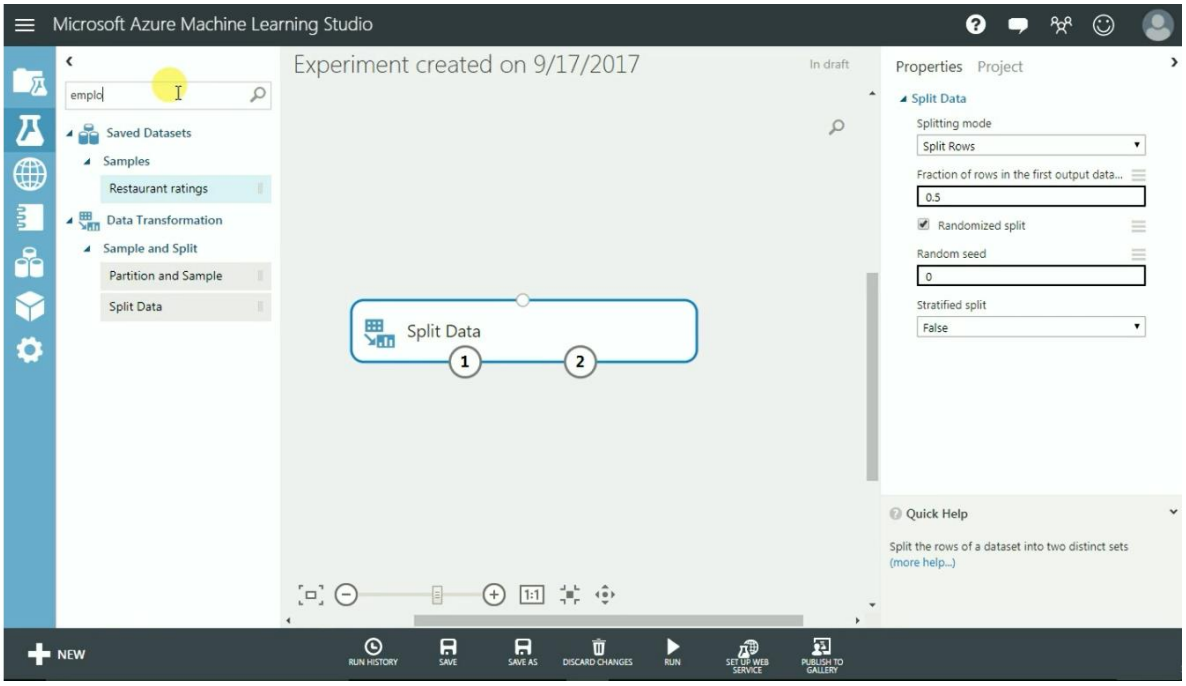


Now you can visualize the result after successful execution

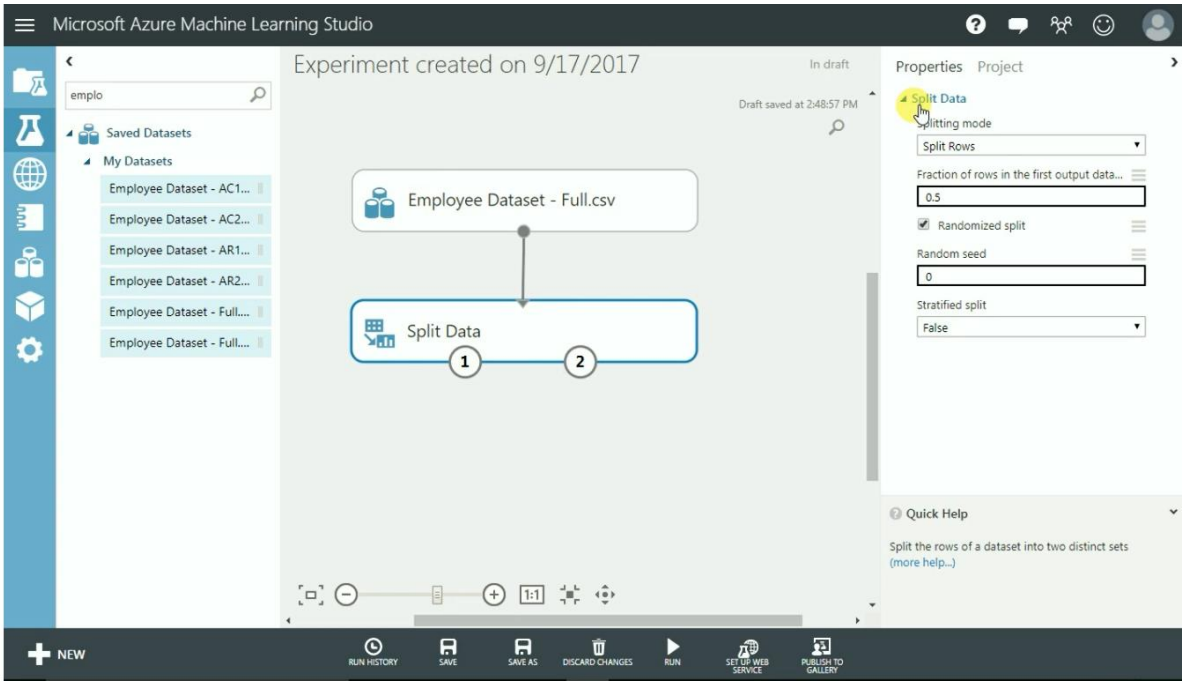


Data Manipulation Using Split Data Component

Search for split data and drop it in the canvas

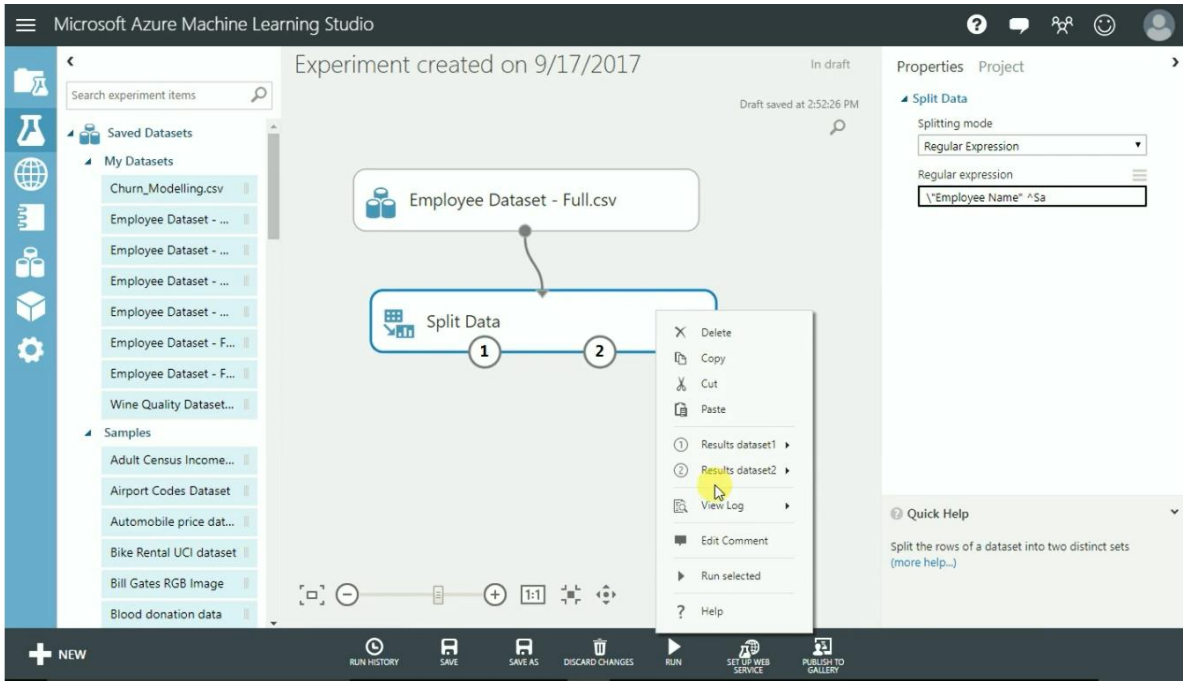


Input existing dataset and connect the nodes

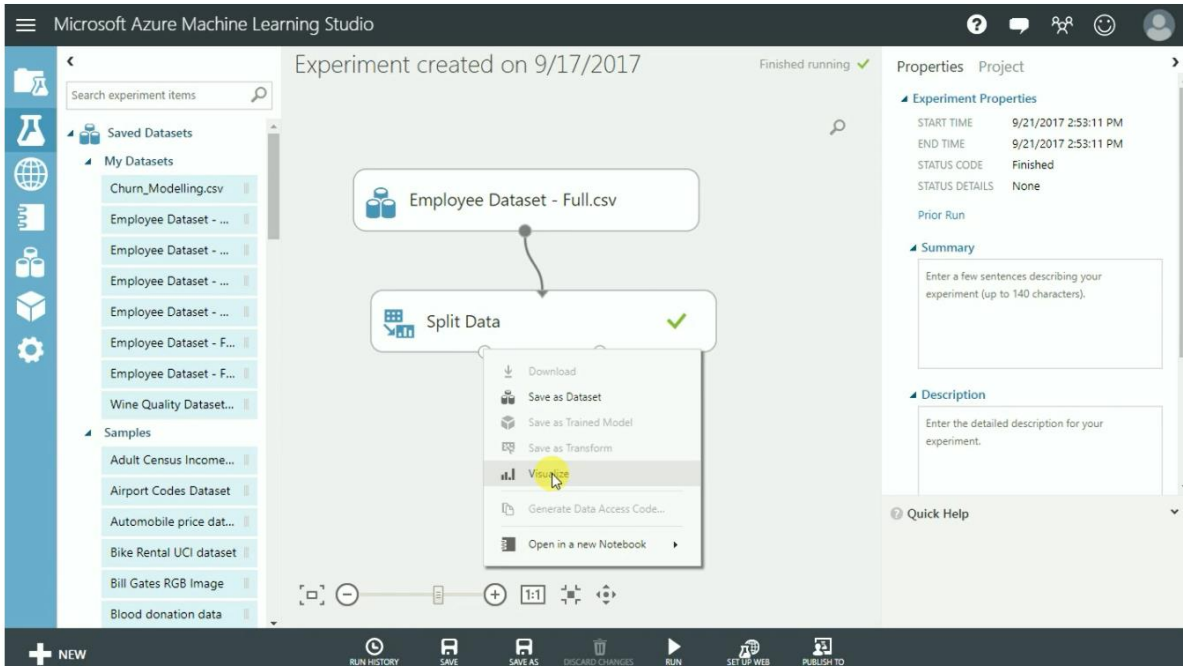


Select parameters as regular expression from drop down and enter

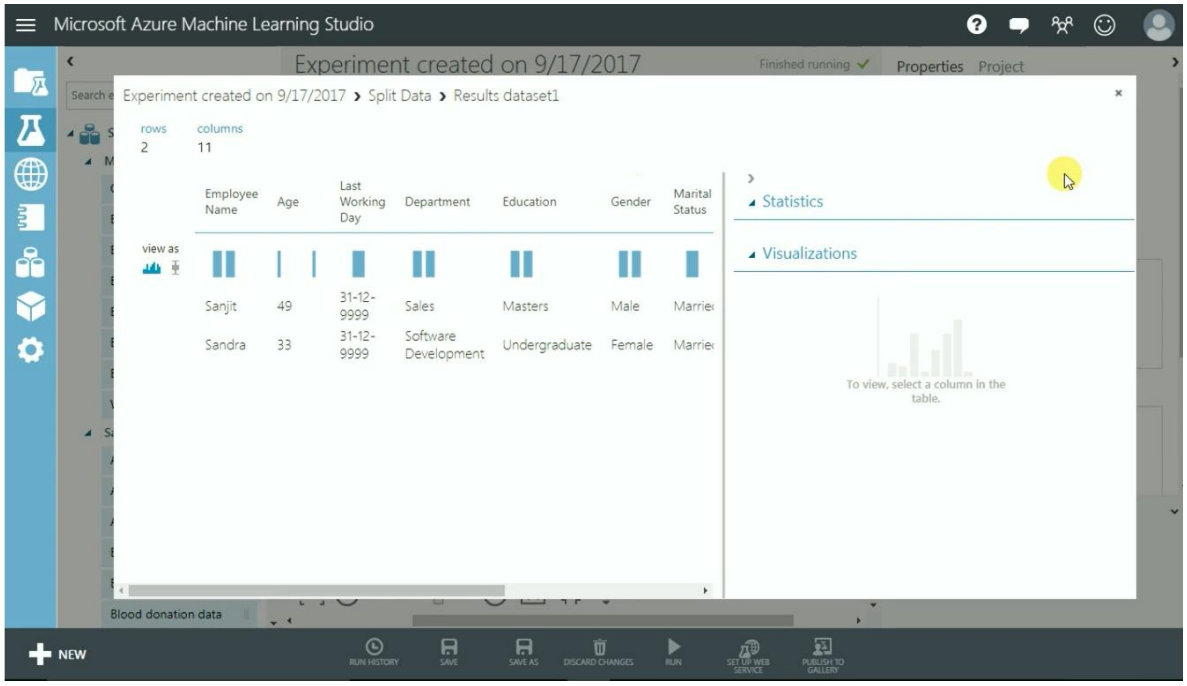
Regular expression with name starting with ‘sa’ and run the module



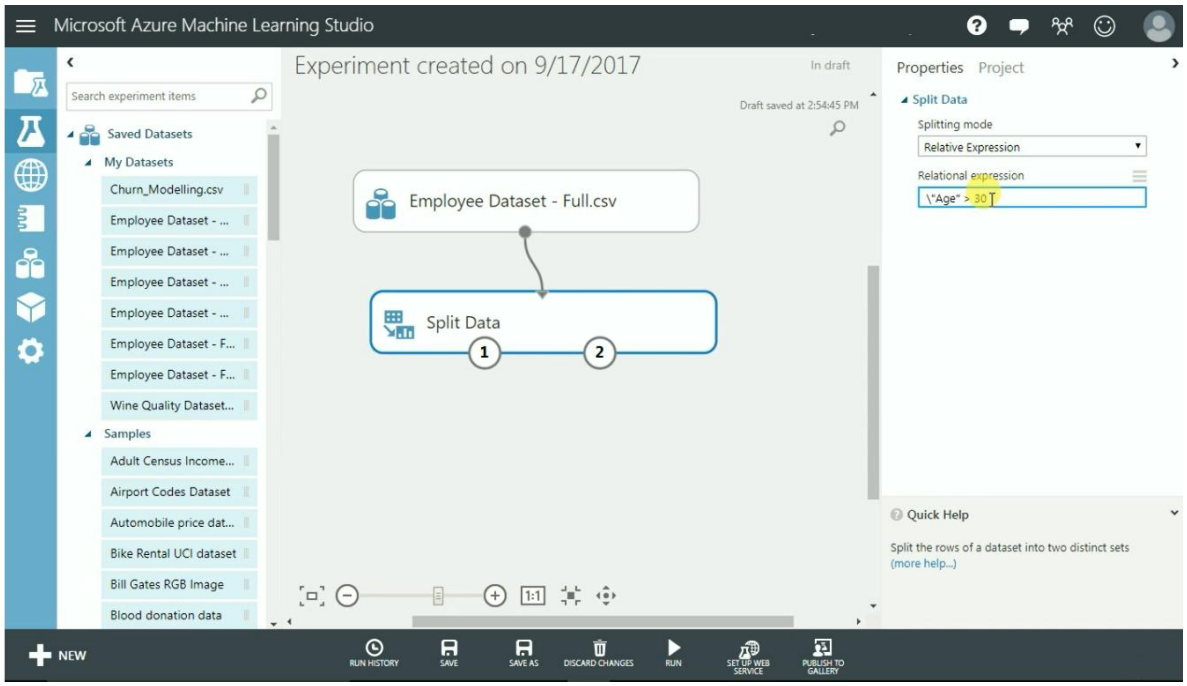
Visualize the output after successful run



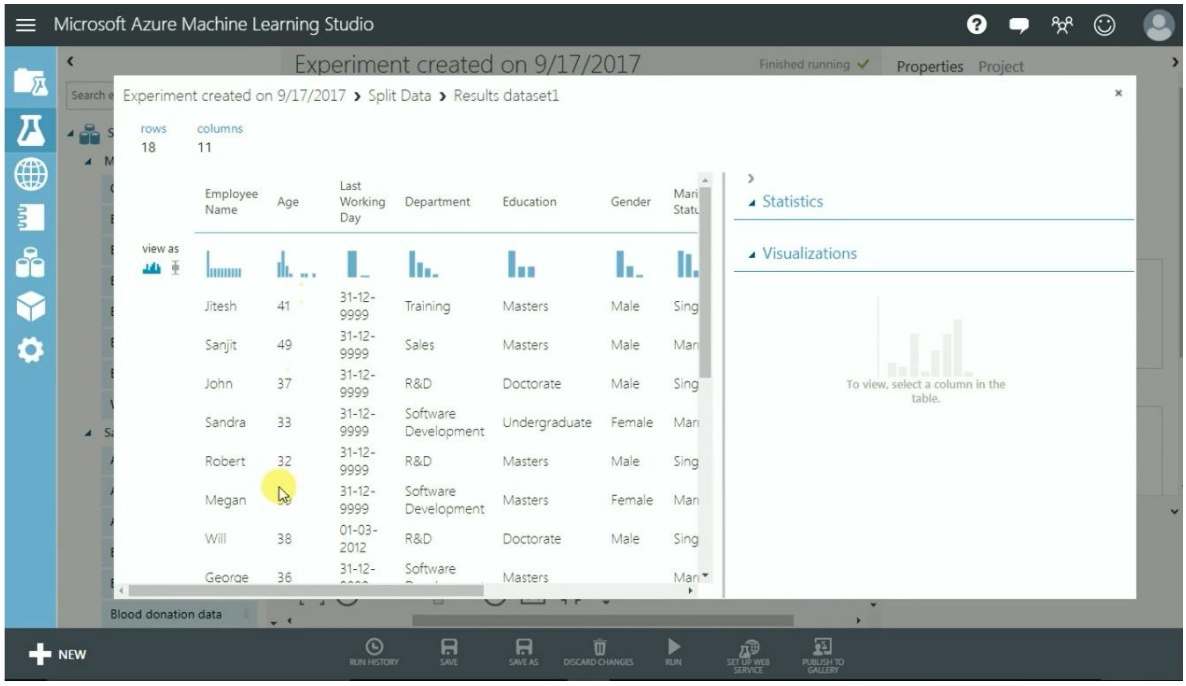
Result successful



Similarly try splitting with relative expression example for employee's age Greater than 30



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



Check node 2 for rest of the dataset