



Hands on Tutorial

k-means Clustering Model

Tutorial #3 Questions:

- What are we trying to accomplish?
 - Build a clustering algorithm to classify wells based on dynamometer card data.
 - Which are similar?
 - Which are dissimilar?
- How can we use Azure Machine Learning to address this question?



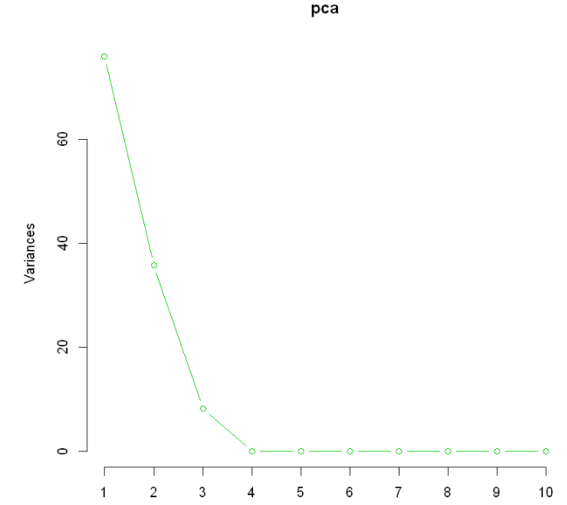
Scenario Data:

- ~3.4 million rows of data
- 263 unique wells
- The data comprises of the following:
 - Date-Time at which the readings were taken
 - Idx - Reading number (1-60): During every up and down motion of the Pump Jack, 60 sets of reading will be taken at various positions
 - X - Position of the Pump Jack
 - Y – Load on the Pump Jack
- We then perform the Principal Component Analysis over this data which was covered during your Exploratory Data Analysis lecture

well	date	idx	x	y
Alabama	8/16/2013 1:55	1	0.00074002	-5327.23698
Alabama	8/16/2013 1:55	2	0.15595834	-4476.758361
Alabama	8/16/2013 1:55	3	1.536036006	-3475.778052
Alabama	8/16/2013 1:55	4	3.779102308	-2636.548613
Alabama	8/16/2013 1:55	5	6.817062356	-1999.545532
Alabama	8/16/2013 1:55	6	10.7311654	-1622.668567
Alabama	8/16/2013 1:55	7	15.5308829	-1432.091513
Alabama	8/16/2013 1:55	8	21.09257393	-1346.745211
Alabama	8/16/2013 1:55	9	27.24284435	-1325.808833

Scenario Data:

- After the PCA was performed we ended up with ~57K rows of data with 3 Principal Components (PC1, PC2, PC3)
- 263 unique wells
- Training data comprises of the following:
 - Date-Time at which the readings were taken
 - PC1
 - PC2
 - PC3

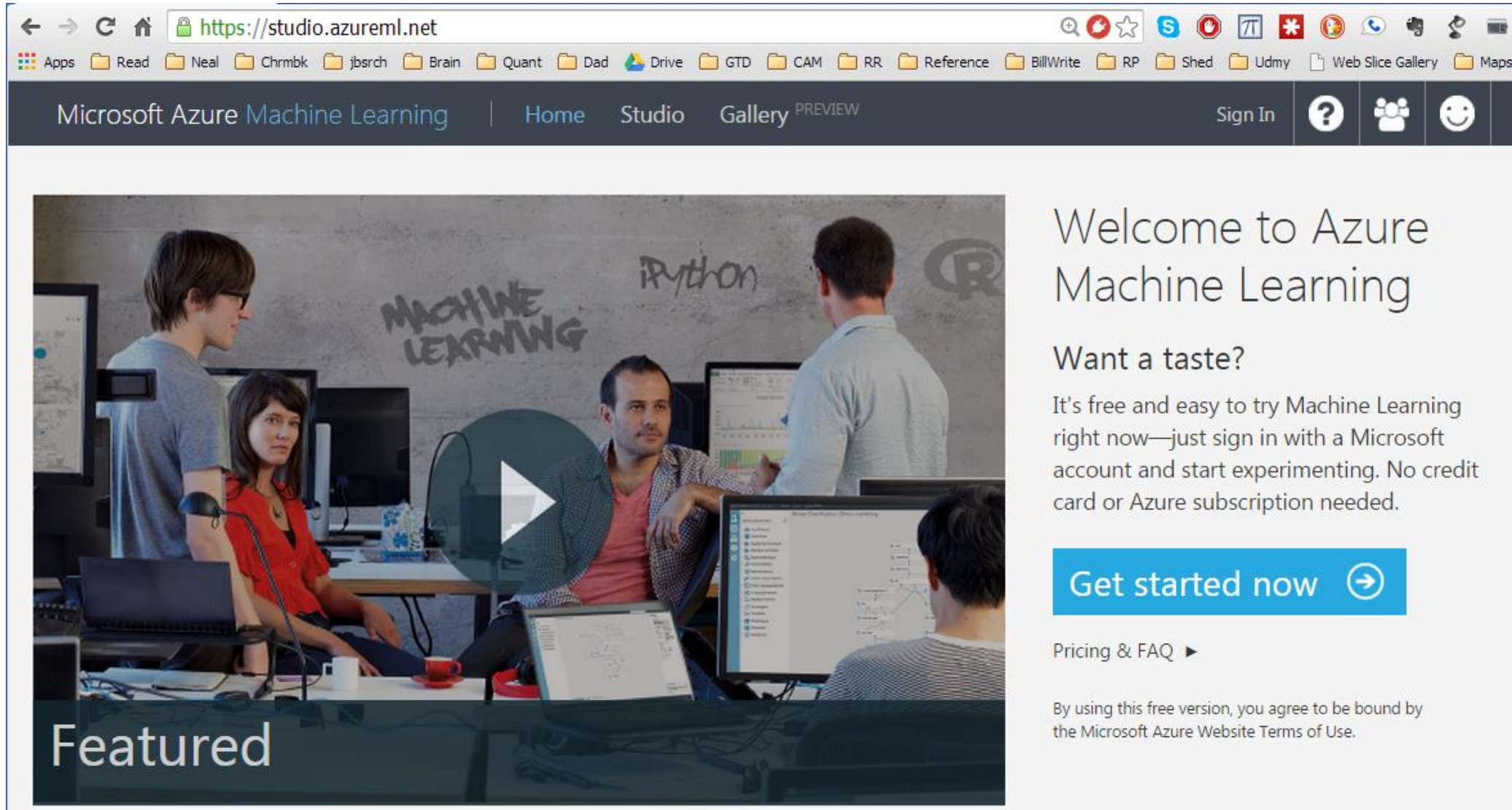


well	date	PC1	PC2	PC3
Alabama	8/16/2013 1:55	-8.243340971	2.918070697	2.023029116
Alabama	8/16/2013 23:49	-8.243340971	2.918070697	2.023029116
Alabama	9/3/2013 0:17	-8.495327766	2.311736767	2.154066737
Alabama	9/6/2013 9:38	-8.148383208	2.731807776	2.341956526
Alabama	9/6/2013 16:13	-8.148383208	2.731807776	2.341956526
Alabama	9/13/2013 3:48	-8.206877815	2.58376763	2.282912974
Alabama	9/13/2013 20:28	-8.206877815	2.58376763	2.282912974
Alabama	11/5/2013 13:17	-8.640142409	2.146508336	2.971266786
Alabama	11/5/2013 17:26	-8.640142409	2.146508336	2.971266786
Alabama	12/12/2013 9:46	-3.139291754	7.206502009	-2.257469677
Alabama	12/12/2013 14:13	-3.139291754	7.206502009	-2.257469677



Step by Step Guide for Building a Clustering Model

Step 1 : Go to <https://studio.azureml.net/>



The screenshot shows the Azure Machine Learning Studio homepage. The browser address bar displays <https://studio.azureml.net/>. The navigation bar includes links for Microsoft Azure Machine Learning, Home, Studio, and Gallery (marked as PREVIEW), along with a Sign In button and user icons. The main content area features a large video player with a play button overlay, showing a group of people working in a machine learning environment. To the right of the video, the text reads "Welcome to Azure Machine Learning" followed by "Want a taste?" and a description: "It's free and easy to try Machine Learning right now—just sign in with a Microsoft account and start experimenting. No credit card or Azure subscription needed." Below this is a prominent blue button labeled "Get started now" with a right arrow icon. Further down, there is a link for "Pricing & FAQ" and a disclaimer: "By using this free version, you agree to be bound by the Microsoft Azure Website Terms of Use." The word "Featured" is visible in the bottom left corner of the video area.

Microsoft Azure Machine Learning | Home Studio Gallery PREVIEW Sign In ? [User Icons]

Welcome to Azure Machine Learning

Want a taste?

It's free and easy to try Machine Learning right now—just sign in with a Microsoft account and start experimenting. No credit card or Azure subscription needed.

[Get started now](#) ➔














[Pricing & FAQ](#) ▶

By using this free version, you agree to be bound by the Microsoft Azure Website Terms of Use.

Featured

Step 2 : Log in to your account

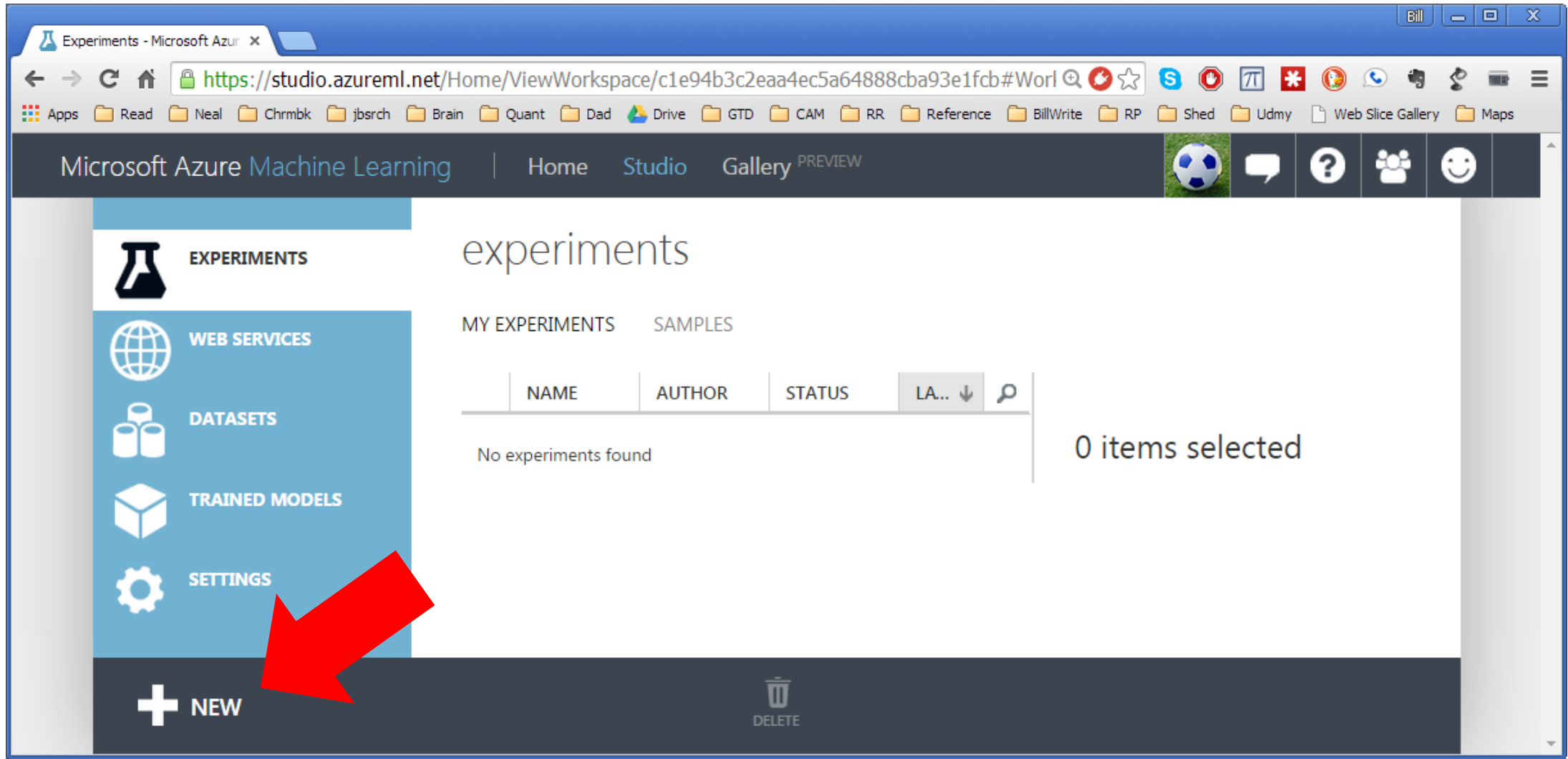
The screenshot shows the Microsoft Azure Machine Learning Studio web interface. The browser address bar displays the URL: <https://studio.azureml.net/Home/ViewWorkspace/c1e94b3c2eaa4ec5a64888cba93e1fcb#World>. The top navigation bar includes 'Microsoft Azure Machine Learning' and links to 'Home', 'Studio', and 'Gallery'. A sidebar on the left contains icons and labels for 'EXPERIMENTS', 'WEB SERVICES', 'DATASETS', 'TRAINED MODELS', and 'SETTINGS'. The main content area is titled 'experiments' and features tabs for 'MY EXPERIMENTS' and 'SAMPLES'. Below these tabs is a table with columns: 'NAME', 'AUTHOR', 'STATUS', and 'LA...'. The table is currently empty, displaying the message 'No experiments found'. To the right of the table, it says '0 items selected'. At the bottom of the interface, there is a '+ NEW' button and a 'DELETE' button with a trash icon.

- ▶  Saved Datasets
- ▶  Data Format Conversions
- ▶  Data Input and Output
- ▶  Data Transformation
- ▶  Feature Selection
- ▶  Machine Learning
- ▶  OpenCV Library Modules
- ▶  Python Language Modules
- ▶  R Language Modules
- ▶  Statistical Functions
- ▶  Text Analytics
- ▶  Deprecated
- ▶  Web Service

Step 3.1 : Create New Experiment

- AML modelling ... a checklist approach
 - ❑ Create new experiment

Step 3.2 : Create a new experiment



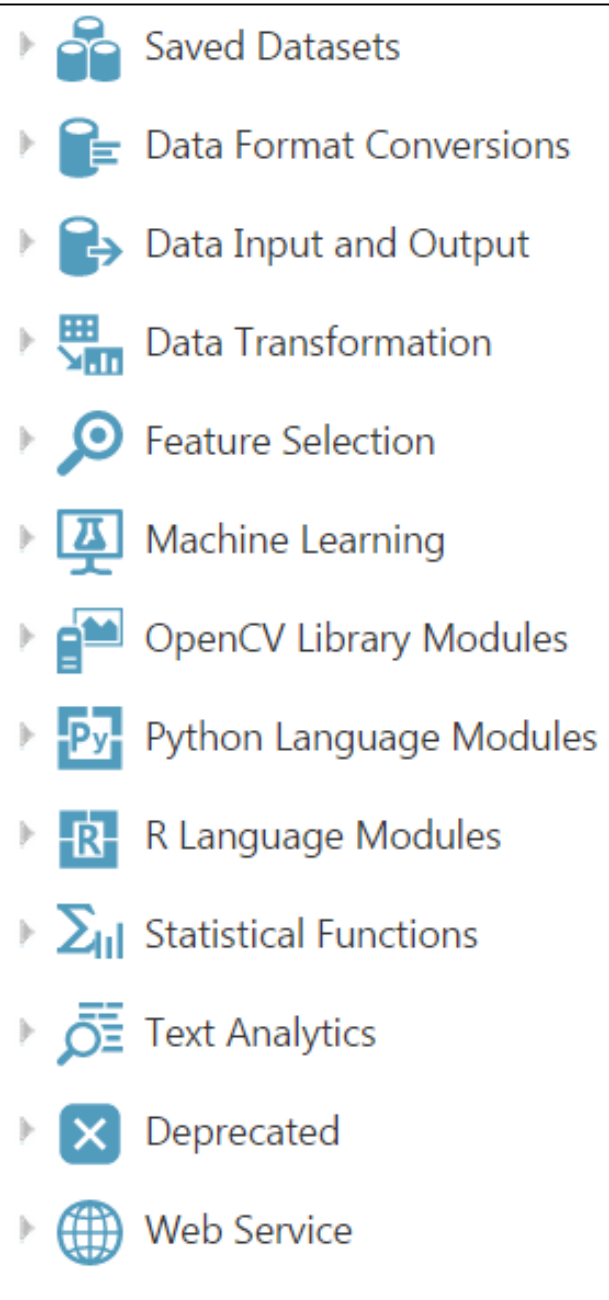
Step 3.3 : Create a new experiment

The screenshot shows the Microsoft Azure Machine Learning web interface. At the top, there's a dark header with the Microsoft Azure Machine Learning logo. Below it, a navigation bar includes 'PROJECTS' and 'EXPERIMENTS' (the latter is highlighted). The main content area is titled 'experiments' and has tabs for 'MY EXPERIMENTS' and 'SAMPLES'. A table with columns 'NAME', 'AUTHOR', and 'STATUS' is visible. On the left, a 'NEW' sidebar lists options: DATASET, MODULE, PROJECT (with a 'PREVIEW' link), EXPERIMENT (highlighted with a red arrow labeled '1'), and NOTEBOOK (with a 'PREVIEW' link). The main area displays 'Microsoft Samples' with a search bar. Below the search bar, there are three cards: 'Blank Experiment' (with a plus icon and a red arrow labeled '2'), 'Experiment Tutorial' (with a green background and a right arrow icon), and 'Sample 1: Download dataset from UCI: Adult 2 class dataset' (with a database cylinder icon and a download arrow).

Step 3.4 : Name your experiment



- Click on the title box at the top that says "Experiment Created on"
- Give the experiment the following title: "Oil & Gas – Dynamometer Clustering"

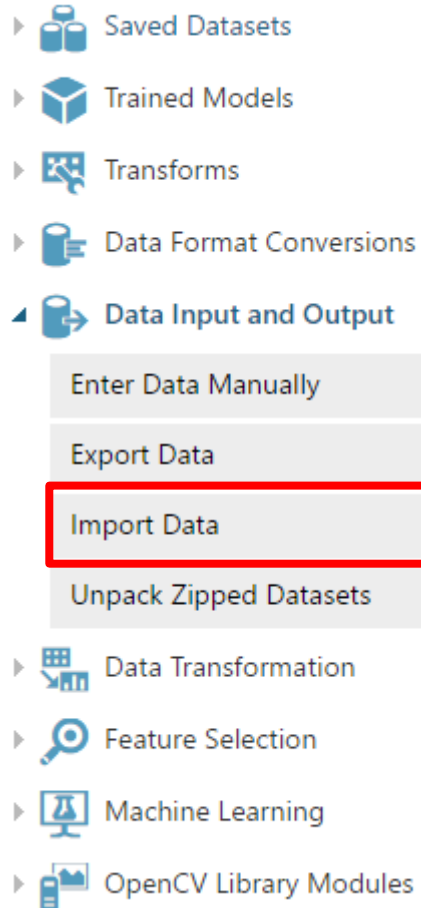


Step 4.1 : Import the dataset

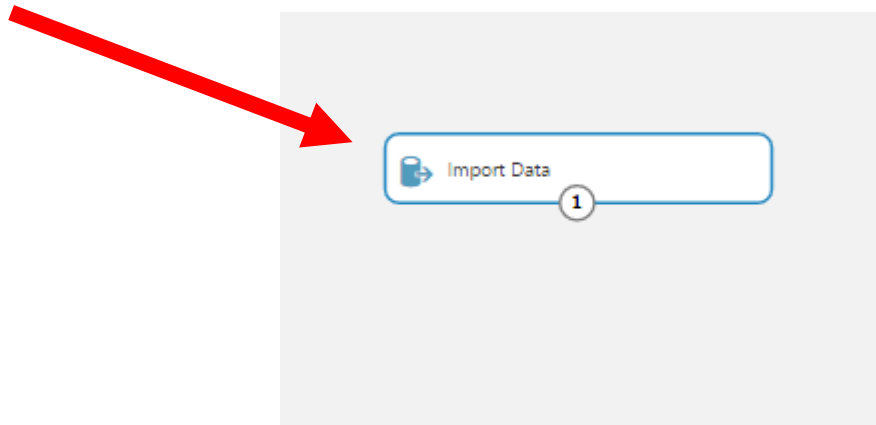
- AML modelling ... a checklist approach
 - ☒ Create new experiment
 - ☐ Import data set

Option 1 -
Import the training dataset
from the Azure Blob Storage. If
option 1 did not work, skip to
option 2.

Step 4.2 : Q? How to import data from Azure Blob?

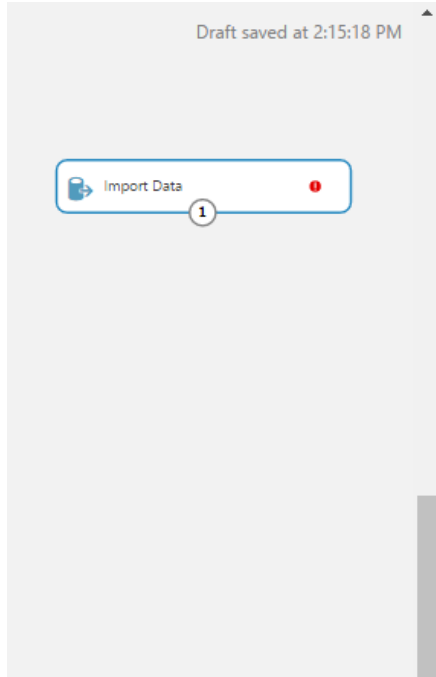


- Open "Data Input and Output" from the navigation pane at the left
- Drag "Import Data" to the canvas
 - "Import Data" loads data from sources such as the Web, Azure SQL, Windows Azure Blob storage, etc



Step 4.3 : Q? How to import data from Azure Blob?

- Machine Learning
 - Click on Import Data
 - Chose "Azure Blob Storage" for Data source
 - Chose "Storage Account" for Authentication type
 - Copy and Paste the following information without the quotes in the Import Data module
 - Account Name – **"nealworkshop"**
 - Account key –
"RER9c7kfM1e67p7p7gl+TbkE5Y6alzURg4PQc9lew+l8O+ZfU58gFjNgBW/WQm0u8N0YZQUG+wla1zfWKxyljA=="
 - Path to container – **"/oilandgas/Dynamometer Card Reading Data PCA.csv"**
 - Check the File has header row check box

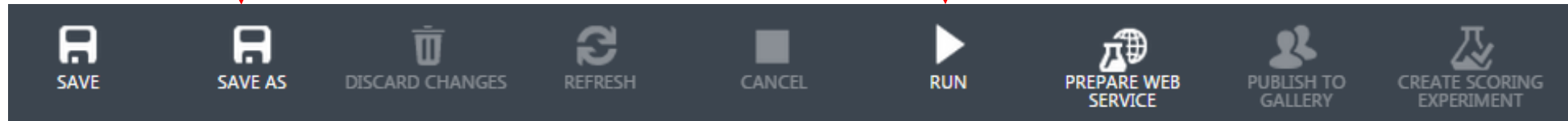


Step 4.4 : Q? How to import data from Azure Blob?

Step 1: Click
"Save"



Step 2: Click "Run"



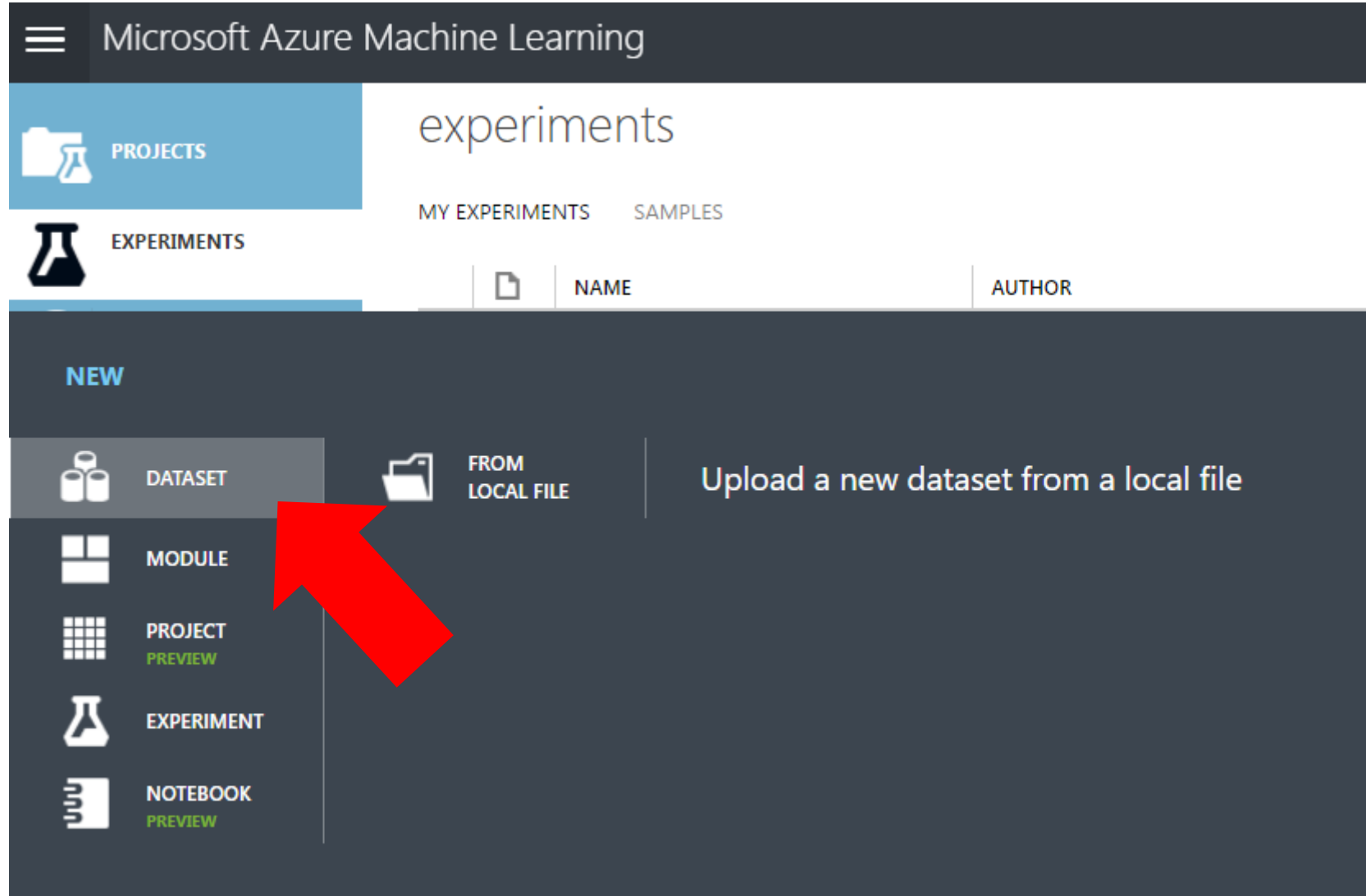
If Option 1 worked, move on to
Step 4.12

Option 2 -
Import the training dataset
from the saved datasets.

Step 4.5 : Import this tutorial's training dataset

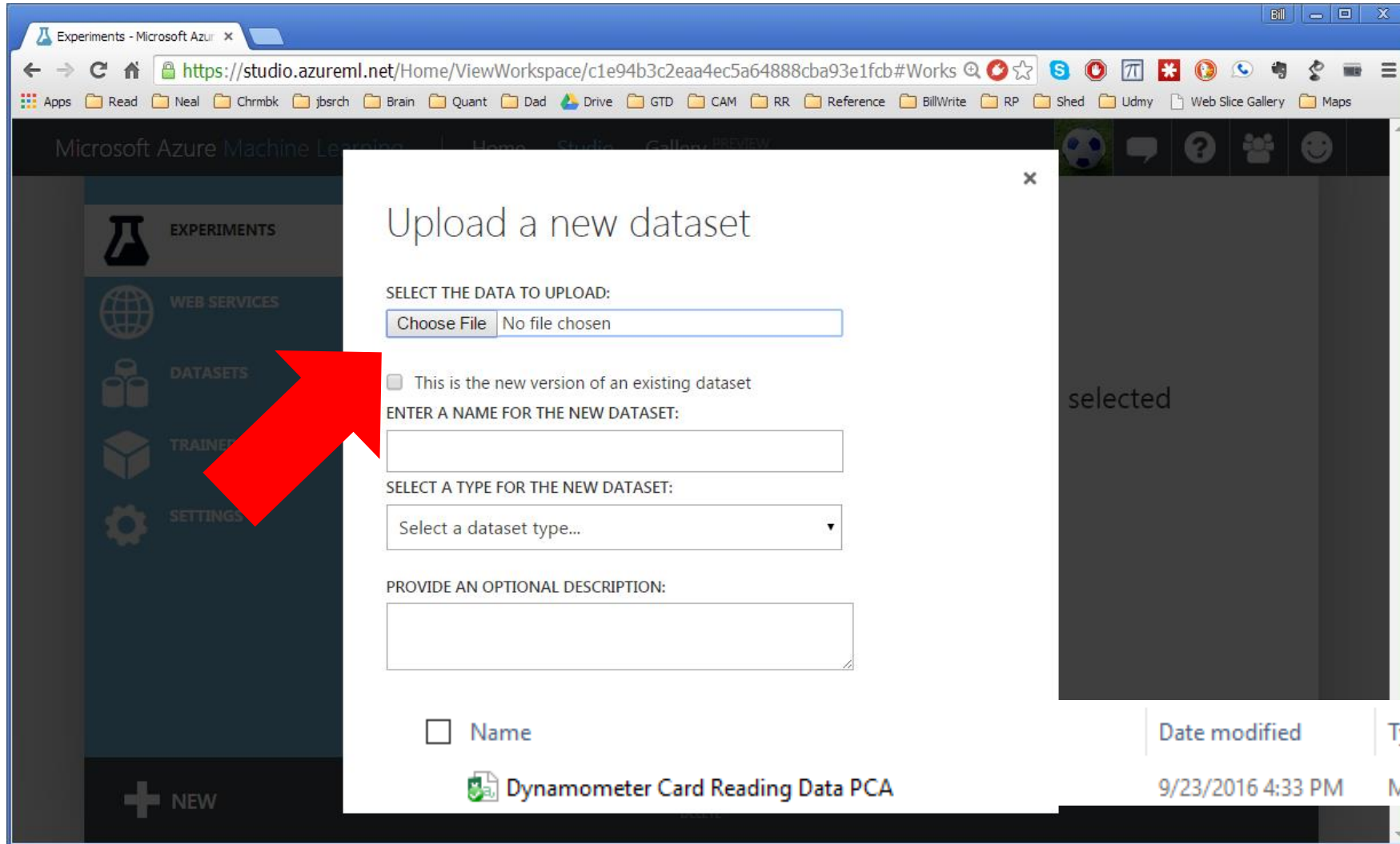
The screenshot shows the Microsoft Azure Machine Learning Studio interface. The browser address bar displays the URL: <https://studio.azureml.net/Home/ViewWorkspace/c1e94b3c2eaa4ec5a64888cba93e1fcb#World>. The page title is "experiments". The left sidebar contains a navigation menu with the following items: EXPERIMENTS (selected), WEB SERVICES, DATASETS, TRAINED MODELS, and SETTINGS. At the bottom of the sidebar is a "+ NEW" button, which is highlighted by a large red arrow. The main content area shows a table with columns: NAME, AUTHOR, STATUS, and LA... (with a dropdown arrow). Below the table, it says "No experiments found". To the right of the table, it says "0 items selected". At the bottom of the main content area, there is a "DELETE" button with a trash icon.

Step 4.6 : Import this tutorial's training dataset



The screenshot displays the Microsoft Azure Machine Learning web interface. At the top, a dark header bar contains the text "Microsoft Azure Machine Learning" next to a hamburger menu icon. Below this, a light blue sidebar on the left features a "PROJECTS" section with a folder and flask icon, and an "EXPERIMENTS" section with a flask icon. The main content area is titled "experiments" and includes tabs for "MY EXPERIMENTS" and "SAMPLES". Below these tabs is a table with columns for a file icon, "NAME", and "AUTHOR". A large dark grey panel at the bottom is titled "NEW" and contains a vertical list of options: "DATASET" (with a flask icon), "MODULE" (with a grid icon), "PROJECT" (with a grid icon and a "PREVIEW" label), "EXPERIMENT" (with a flask icon), and "NOTEBOOK" (with a notepad icon and a "PREVIEW" label). A large red arrow points from the "DATASET" option towards the right. To the right of the "NEW" panel, there is a section titled "FROM LOCAL FILE" with a folder icon and the text "Upload a new dataset from a local file".

Step 4.7 : Import this tutorial's training dataset




The screenshot shows the Microsoft Azure Machine Learning Studio interface. A dialog box titled "Upload a new dataset" is open, overlaying the main workspace. The dialog box contains the following fields and options:

- SELECT THE DATA TO UPLOAD:** A button labeled "Choose File" and a text field showing "No file chosen".
- ☐ This is the new version of an existing dataset
- ENTER A NAME FOR THE NEW DATASET:** A text input field.
- SELECT A TYPE FOR THE NEW DATASET:** A dropdown menu with the text "Select a dataset type..." and a downward arrow.
- PROVIDE AN OPTIONAL DESCRIPTION:** A text input field.

A large red arrow points from the left sidebar towards the "Choose File" button in the dialog box.

In the background, the "DATASETS" section of the sidebar is visible. Below the dialog box, a table lists the datasets in the workspace:

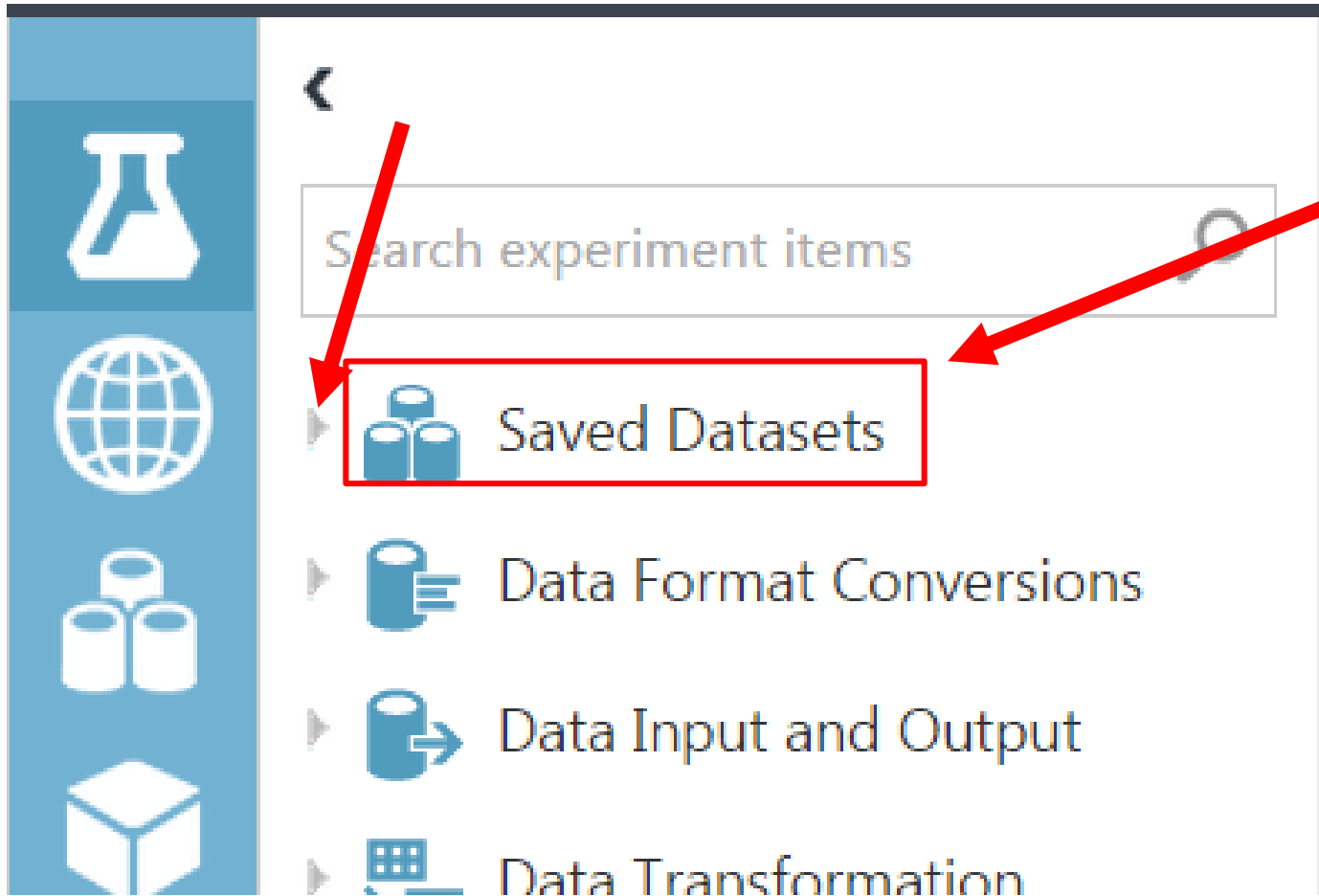
Name	Date modified	Type	Size
 Dynamometer Card Reading Data PCA	9/23/2016 4:33 PM	Microsoft Excel C...	4,620 KB

Step 4.8 : Import this tutorial's training dataset

The screenshot displays the Microsoft Azure Machine Learning Studio interface. The left sidebar contains navigation options: PROJECTS, EXPERIMENTS, WEB SERVICES, NOTEBOOKS, DATASETS, TRAINED MODELS, and SETTINGS. The main area is titled 'experiments' and shows a table of experiments. The table has columns for NAME, AUTHOR, and STATUS. The first experiment, 'Oil & Gas - Dynamometer Clustering', is highlighted. Below the table, a workflow diagram is visible, showing steps like 'K Means Clustering', 'Sweep Clustering', and 'Assign Data to Clusters'. A green arrow points to the 'Sample 6: Train, Test, Evaluate for Regression: Auto Imports Dataset' experiment. A status bar at the bottom indicates 'Upload of the dataset 'Dynamometer Card Reading Data PCA.csv' has completed.' with an 'OK' button.

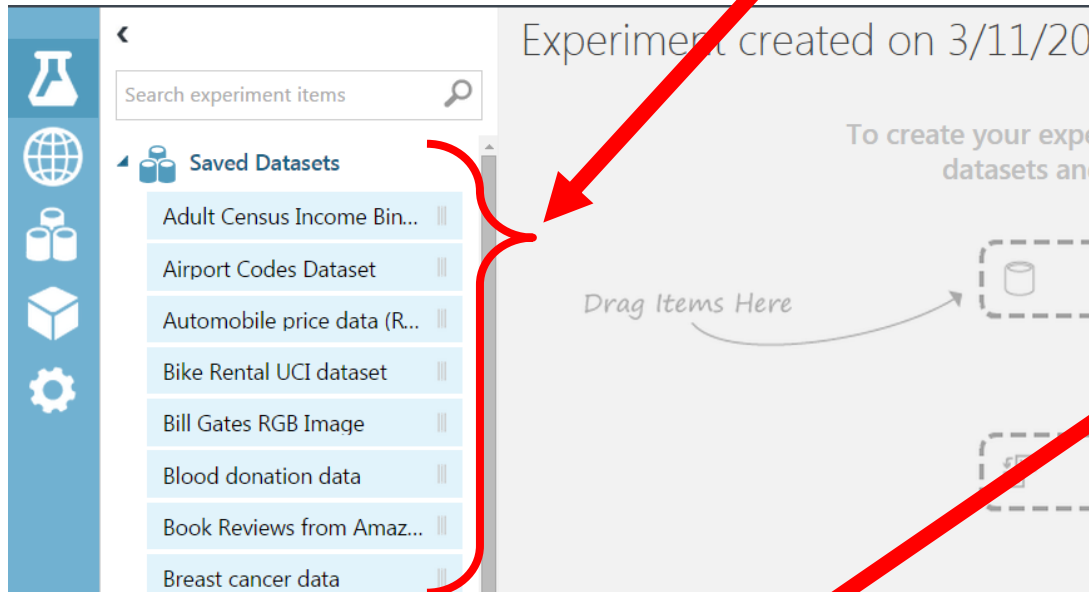
NAME	AUTHOR	STATUS
Oil & Gas - Dynamometer Clustering	acha_mallaya	Draft
OK Cluster, Classify, Regress Dynamo	acha_mallaya	Draft
Oil & Gas - Dynamometer Regression	acha_mallaya	Draft
Oil & Gas - Dynamometer Classification	acha_mallaya	Draft
Oil & Gas - Binary Classification Model to predict if the compressor is going ...	acha_mallaya	Draft
Oil & Gas - Linear Regression Model to predict the Remaining Useful Life	acha_mallaya	Draft
OK Training - Tank Level Forecasting regression	sailaja.karthik	Finishe
OK Predictive Maintenance: Step 1 of 3, data preparation and feature engine...	eric.hullander	Draft
OK - K Means Clustering - Brine Analysis	sailaja.karthik	Draft
Oil & Gas - Binary Classification Model to predict if the compressor is going ...	sailaja.karthik	Failed
Oil & Gas - Linear Regression Model to predict the Remaining Useful Life [Pr...	acha_mallaya	Draft
Binary Classification: Breast cancer detection	Microsoft	Draft
OK Predictive Maintenance: Step 2B of 3, train and evaluate binary classifica...	Microsoft	Draft
e - Linear Regress... el to predict the R... Useful Life -	Finishe
Tracking Success v... ization	...	Finishe
Tracking Success v... ization	...	Finishe
Sample 6: Train, Test, Evaluate for Regression: Auto Imports Dataset	Microsoft	Finishe

Step 4.9 : Open "Saved Datasets"

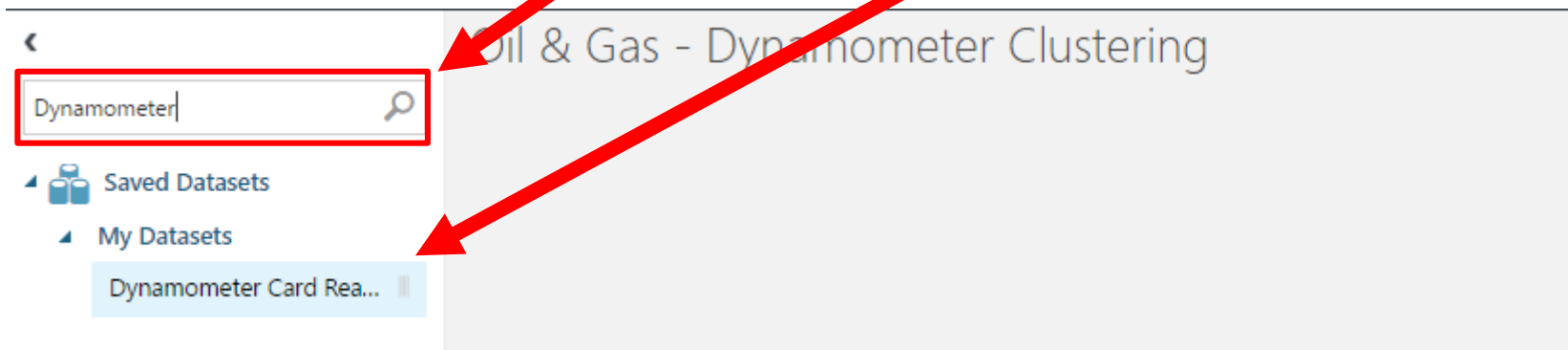


- By clicking on the triangle at the left of "Saved Datasets"

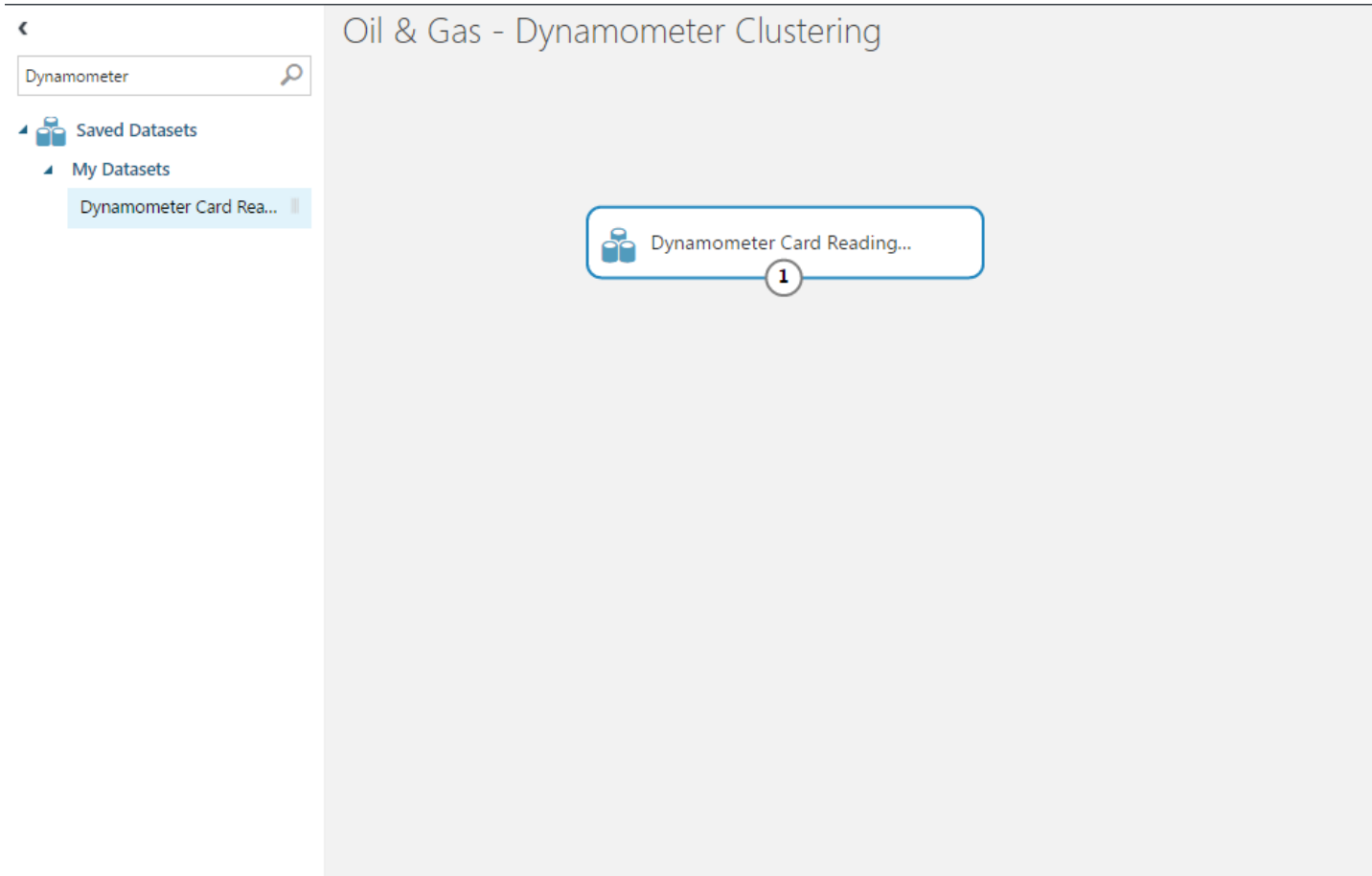
Step 4.10 : Take a second to notice the MANY datasets



- To filter to the data set for this tutorial ...
- Type "Dynamometer Card Reading" in the "Search experiment items" dialog box
- The data set list will reduce to our data set for this tutorial

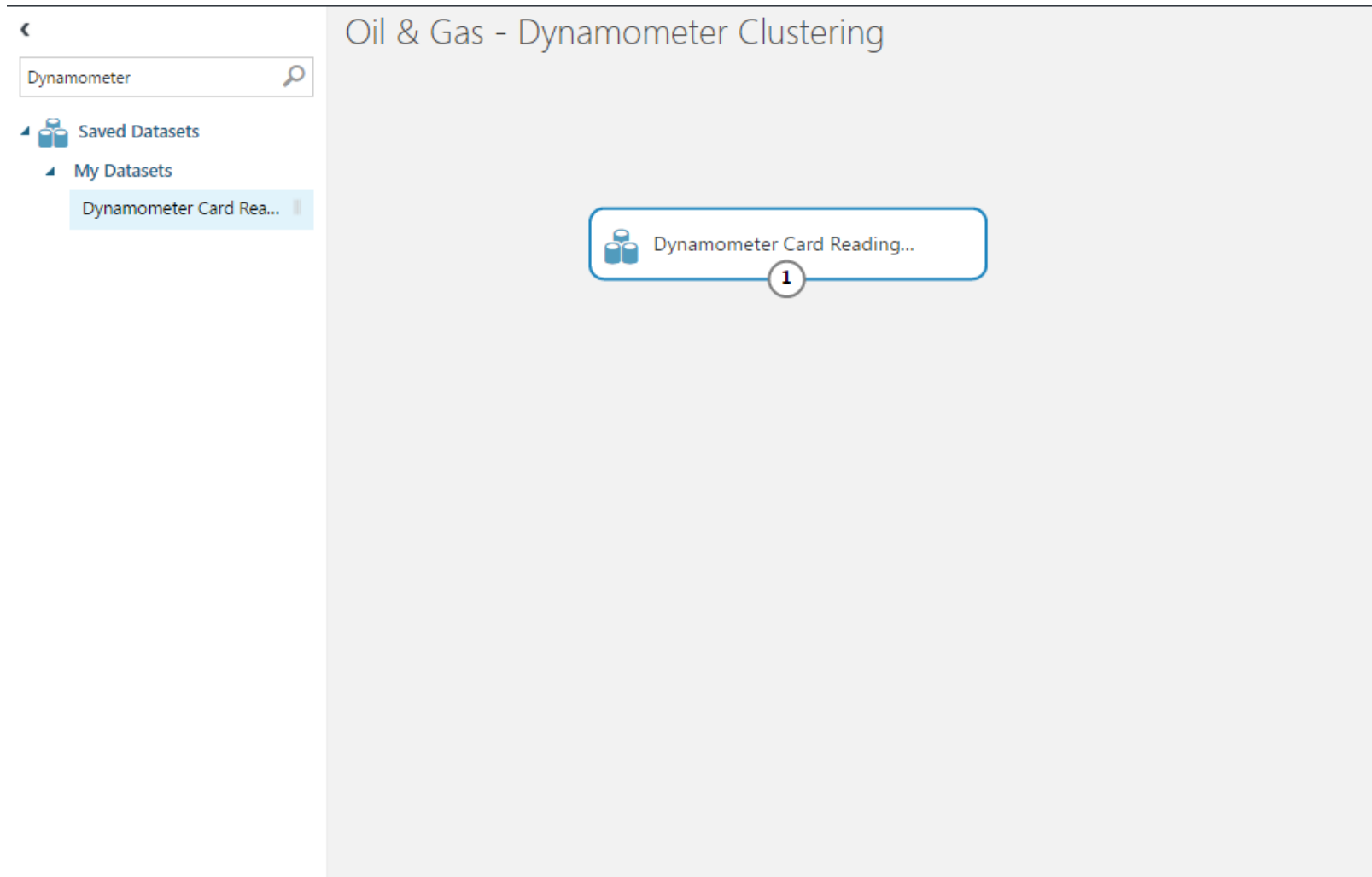


Step 4.11 : Drag the data set to the experiment



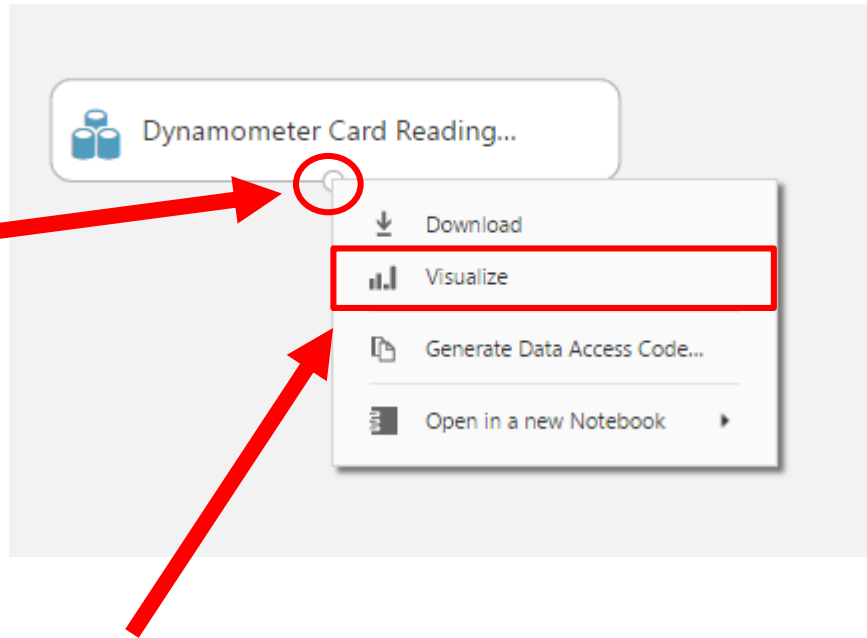
- *Note* when you drag the first element of your model to the canvas ... all the guides disappear
- Now, where are all the tools that were at the left Azure Machine Learning?
- They are still there, ... but we need to un-filter to see them

Step 4.12 : Admire your data set living in your experiment



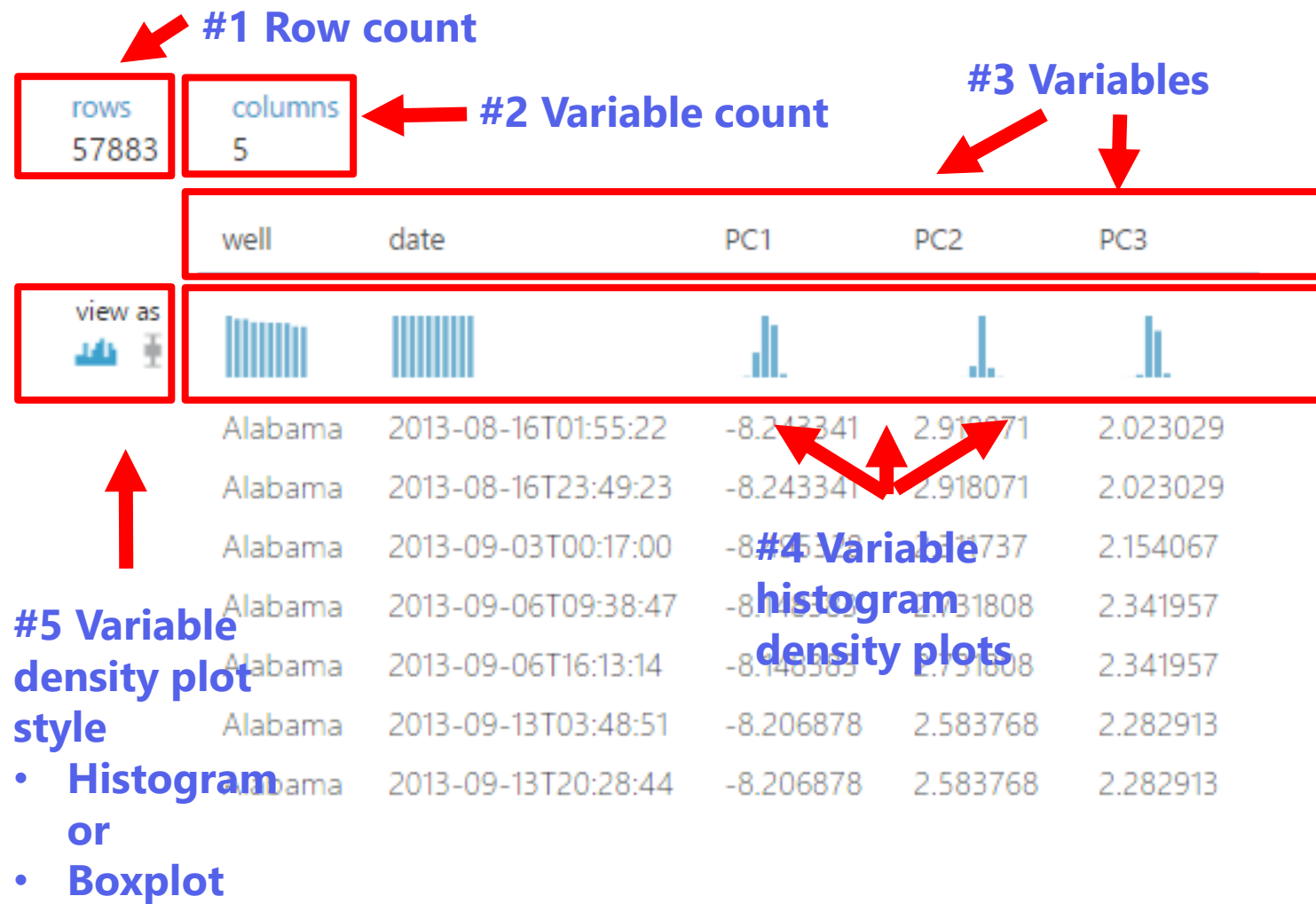
Step 4.13 : Visualize the data set

Right click on the
bottom-middle
circle of the
module
















Then click "Visualize"

Step 4.14 : Now look at the data for 4 attributes

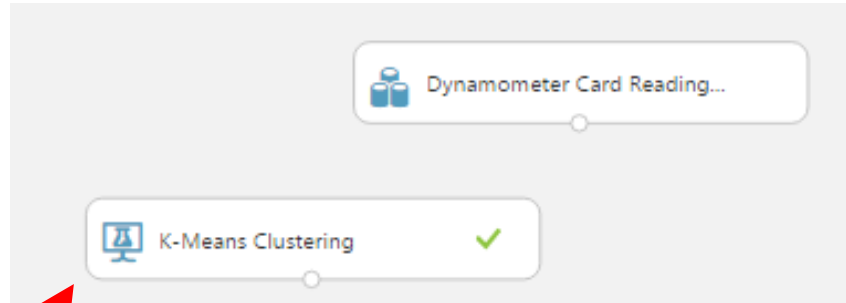
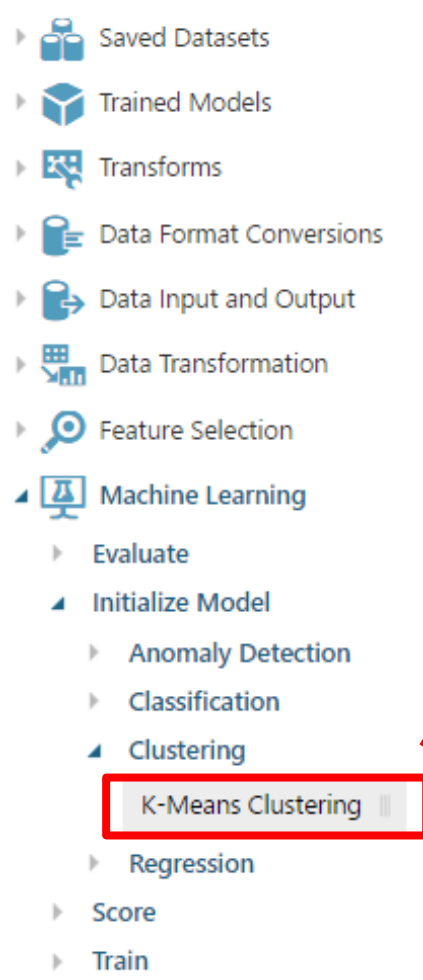


Step 5.1 : AML Modeling Components

- ▶  Saved Datasets
- ▶  Data Format Conversions
- ▶  Data Input and Output
- ▶  Data Transformation
- ▶  Feature Selection
- ▶  Machine Learning
- ▶  OpenCV Library Modules
- ▶  Python Language Modules
- ▶  R Language Modules
- ▶  Statistical Functions
- ▶  Text Analytics
- ▶  Deprecated
- ▶  Web Service

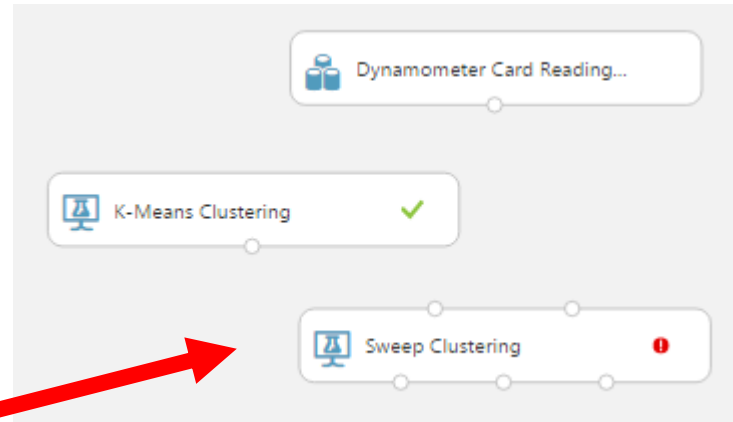
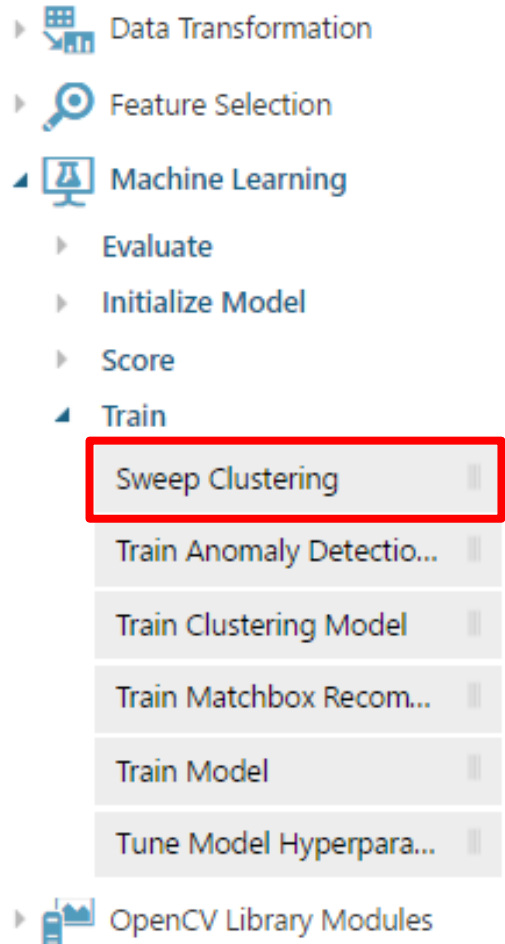
- AML modelling ... a checklist approach
 - ☒ Initial data set
 - ☒ Create new experiment
 - ☒ Import data set
 - ☐ Machine Learning

Step 5.2 : Q? How to develop the Machine Learning Algorithm?



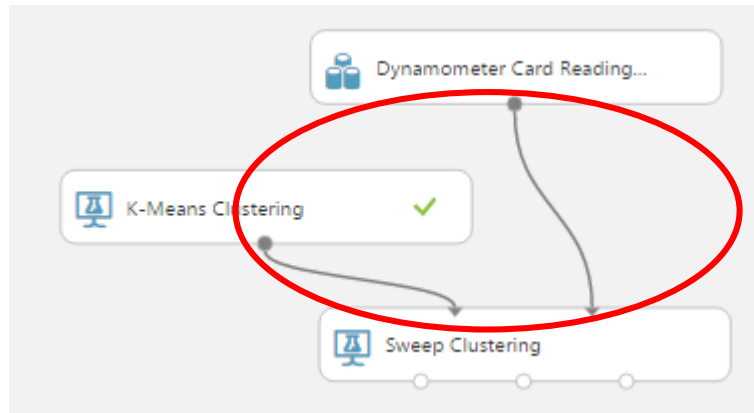
- Open "Machine Learning" -> "Initialize Model" -> "Clustering" from the navigation pane at the left
- Drag "K-Means Clustering" to the canvas

Step 5.3 : Q? How to develop the Machine Learning Algorithm?



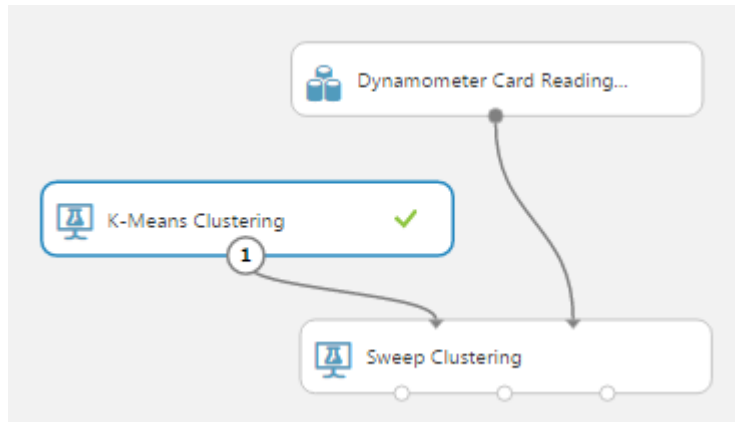
- Open "Machine Learning" and then "Train" from the navigation pane at the left
- Drag "Sweep Clustering" module to the canvas
 - "Train Clustering Model" performs a sweep on the clustering model to determine the optimum parameter settings

Step 5.4 : Q? How to develop the Machine Learning Algorithm?



- Next, click and hold on the bottom middle circle of your "K-Means Clustering" module and drag the line to the top left circle of the "Sweep Clustering" module.
- Similarly, click and hold on the bottom left circle of your "Dataset" module and drag the line to the top right circle of the "Sweep Clustering" module.

Step 5.5: Q? How to develop the Machine Learning Algorithm?



- Machine Learning
 - Initialize parameters for K-Means Clustering algorithm
 - Click on “K-Means Clustering” module on your canvas, and make sure it has a blue outline
 - A properties window will open on extreme right

▲ K-Means Clustering

Create trainer mode
Parameter Range ▼

Range for Number of Cen...
☒ Use Range Builder
Parameter Range : 2 - 66
[Slider bar]

Number of points : 3
☐ Log Scale

Initialization for sweep
K-Means++ ▼

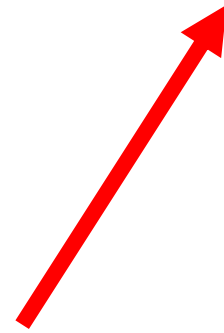
Random number seed
1234

Number of seeds to sweep
1

Metric
Euclidean ▼

Iterations
100

Assign Label Mode
Ignore label column ▼



Step 5.6: Q? How to develop the Machine Learning Algorithm?

- Machine Learning
 - ~~Click on K-Means Module and a properties window will open on extreme right~~
 - Initialize parameters

Default Values

K-Means Clustering

Create trainer mode

Single Parameter

Number of Centroids

2

Initialization

K-Means++

Random number seed

Metric

Euclidean

Iterations

100

Assign Label Mode

Ignore label column

Number of Clusters needed

Initialization – Algorithm used for initial cluster configuration

Metric – Function used to measure distance between cluster vectors

Iterations – Number of times the allocation is retested

Step 5.7 : Q? How to develop the Machine Learning Algorithm

▲ K-Means Clustering

Create trainer mode

Single Parameter ▼

Number of Centroids

2

Initialization

K-Means++ ▼

Random number seed

Metric

Euclidean ▼

Iterations

100

Assign Label Mode

Ignore label column ▼



Assign Values

▲ K-Means Clustering

Create trainer mode

Parameter Range ▼

Range for Number of Cen...

☒ Use Range Builder

Parameter Range : 2 - 66

Number of points : 3

☐ Log Scale

Initialization for sweep

K-Means++ ▼

Random number seed

1234

Number of seeds to sweep

1

Metric

Euclidean ▼

Iterations

100

Assign Label Mode

Ignore label column ▼

- Machine Learning
 - ~~= Click on K-Means Module and a properties window will open on extreme right~~
 - Initialize parameters

Starting with 4 Clusters and rest as default options.

Step 5.8 : Q? How to develop the Machine Learning Algorithm



◀ Sweep Clustering

Metric for measuring clust...
Dunn

Specify parameter sweeping m...
Random sweep

Maximum number of runs...
5

Random seed
1234

Column Set

Selected columns:

All columns

Exclude column names:

well,date

Launch column selector

☒ Check for Append or ...

- Machine Learning
 - Click on Sweep Cluster
 - Set Metric for measuring clustering to "Dunn"
 - Set the Random seed to "1234"
 - Then click on "Launch column selector"

Step 5.9 : Q? How to develop the Machine Learning Algorithm

Select columns

BY NAME
WITH RULES

☐ Allow duplicates and preserve column order in selection

Begin With
ALL COLUMNS NO COLUMNS

Include column indices Enter column indices

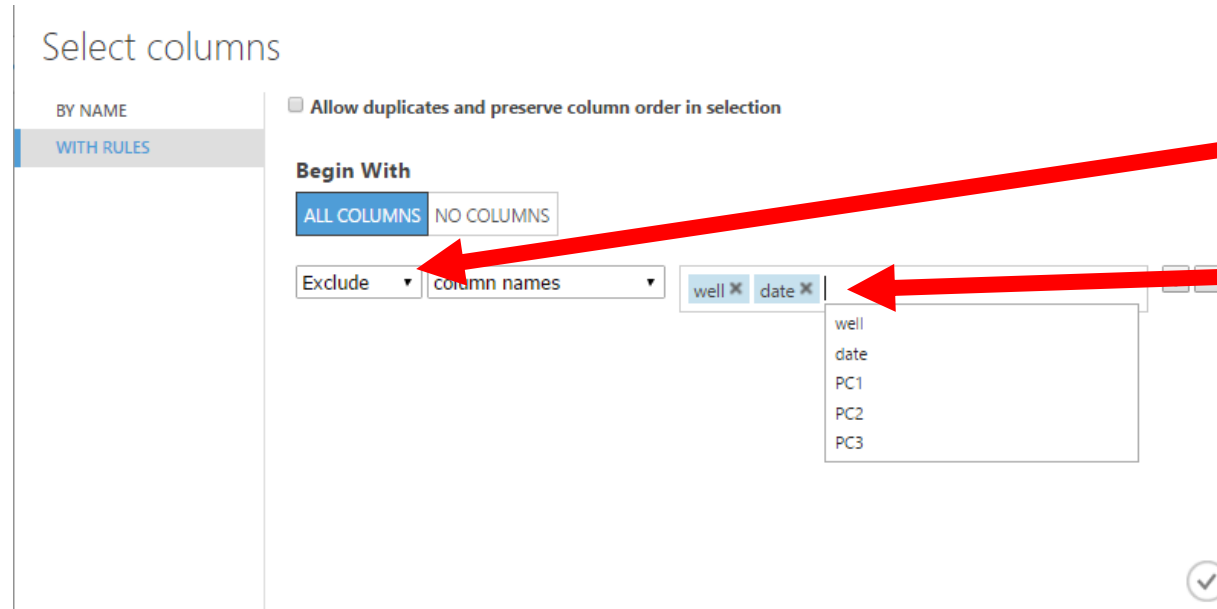
+

-

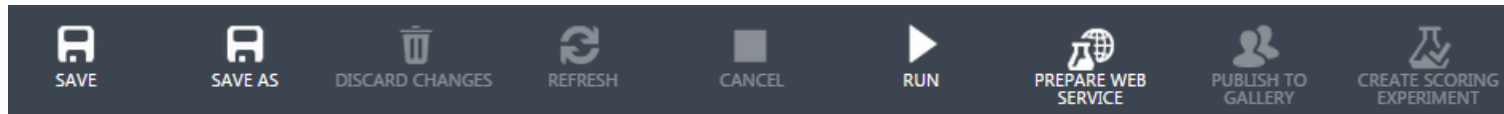
✓

- Machine Learning
 - Select "With Rules"
 - Select "All Columns" from list
- Begin with***

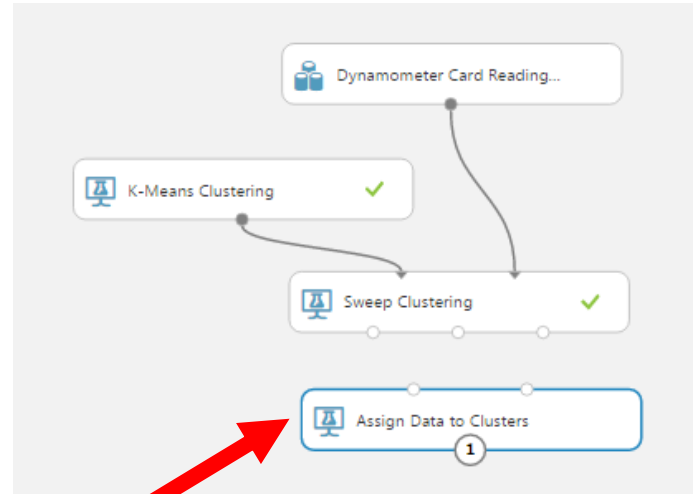
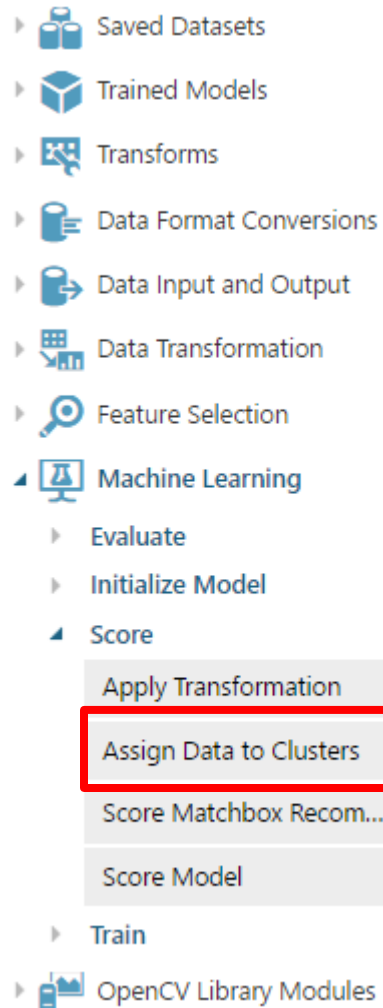
Step 5.10 : Q? How to develop the Machine Learning Algorithm



- Machine Learning
 - Select Exclude from the dropdown
 - Chose the following columns:
 - well
 - date
 - Then click on Check sign to close window
 - Click "Run"

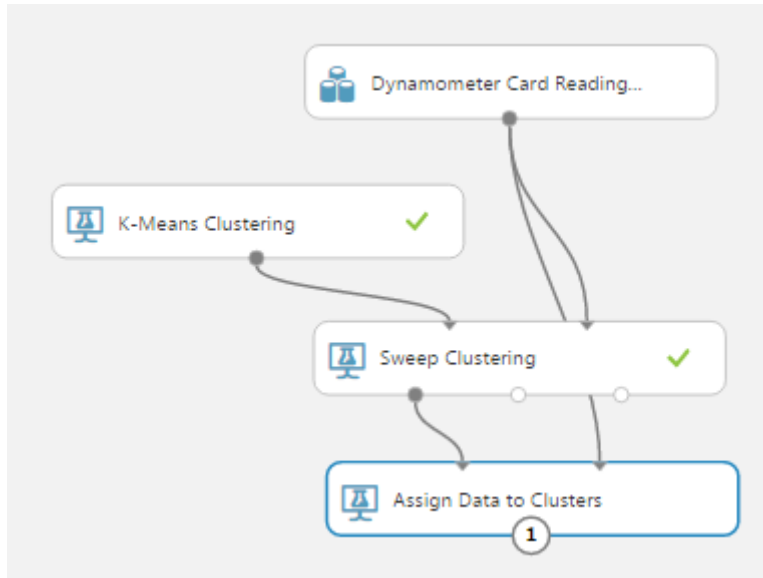


Step 5.11 : Q? How to develop the Machine Learning Algorithm?



- Open "Machine Learning" and then "Score" from the navigation pane at the left
- Drag "Assign Data to Clusters" module to the canvas
 - "Assign Data to Clusters" assigns data to clusters using an existing trained clustering model

Step 5.12 : Q? How to develop the Machine Learning Algorithm?

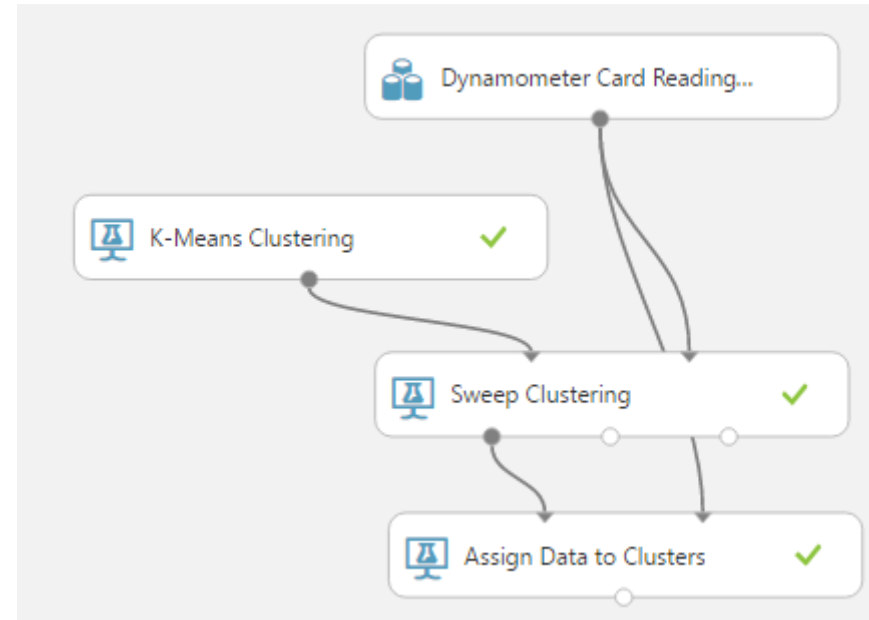


- Next, click and hold on the bottom left circle of your "Sweep Clustering" module and drag the line to the top left circle of the "Assign Data to Clusters" module.
- Similarly, click and hold on the bottom middle circle of your "Dataset" module and drag the line to the top right circle of the "Assign Data to Clusters" module.

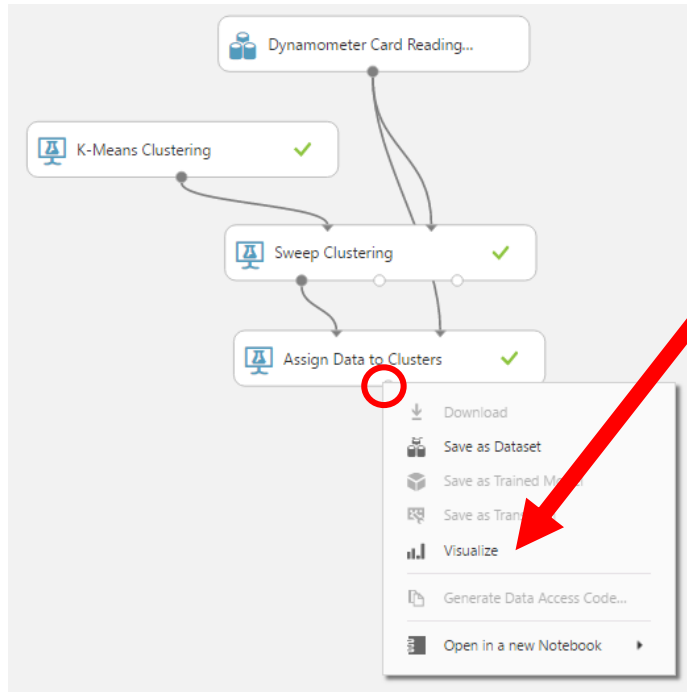
Step 5.13 : Run the Experiment



Click "Run"

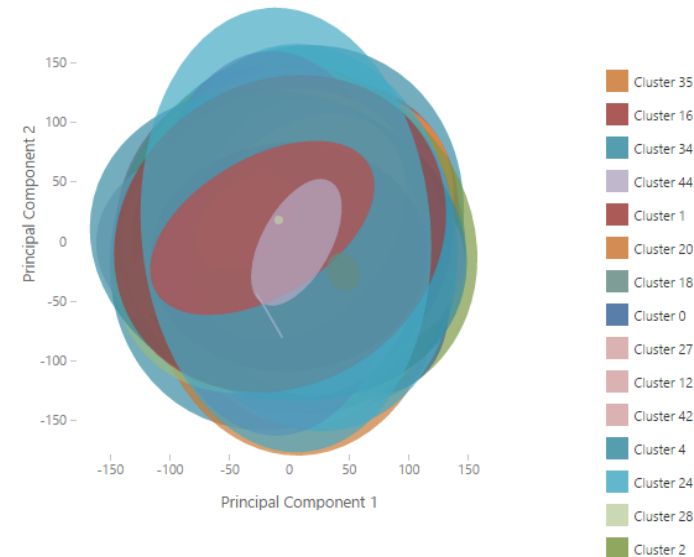


Step 5.14 : Visualize the Results

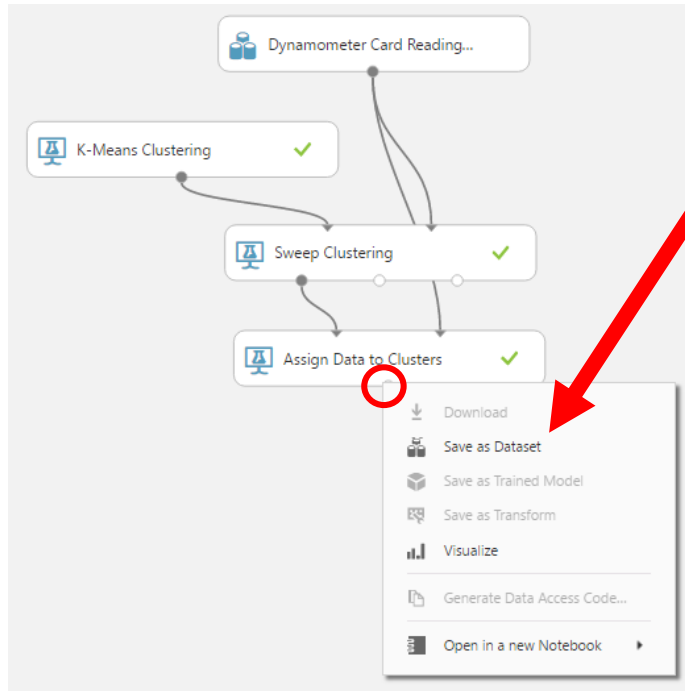


- The Clustering algorithm has assigned a cluster across each item given by variable "Assignments"
- Right click on the bottom-middle circle of "Assign Data to Clusters" module and select "Visualize"

Oil & Gas - Dynamometer Clustering > Assign Data to Clusters > Results dataset



Step 5.15 : Save the Result as a Dataset



- Right click on the bottom-middle circle of "Assign Data to Clusters" module and select "Save as Dataset"
- A pop-up will appear. Call the new dataset "Dynamometer Clusters"
- Click on the check mark

The screenshot shows a pop-up dialog titled "Save output as a new dataset". It has a close button (X) in the top right corner. The dialog contains a checkbox labeled "This is the new version of an existing dataset" which is checked. Below this is a text input field labeled "Enter a name for the new dataset:" with the text "Dynamometer Clusters" entered. Below that is a text area labeled "Provide an optional description:". At the bottom right, there is a checkmark button. A red arrow points to this checkmark button.

Step 5.16 : Save the Result as a Dataset

The screenshot displays the Microsoft Azure Machine Learning Studio interface for an experiment titled "Oil & Gas - Dynamometer Clustering". The interface includes a left-hand navigation pane with categories like Saved Datasets, Trained Models, Transforms, Data Format Conversions, Data Input and Output, Data Transformation, Feature Selection, Machine Learning, Train, and Statistical Functions. The Machine Learning section is expanded, showing options like Evaluate, Initialize Model, Score, Apply Transformation, Assign Data to Clust..., Score Matchbox Rec..., and Score Model. The main workspace shows a workflow diagram with four nodes: "Dynamometer Card Reading...", "K-Means Clustering", "Sweep Clustering", and "Assign Data to Clusters". The "K-Means Clustering", "Sweep Clustering", and "Assign Data to Clusters" nodes are marked with green checkmarks, indicating they have been executed successfully. Below the workflow diagram, there are six large green arrows pointing downwards, suggesting the next steps in the process. The right-hand pane shows the "Properties" tab with "Experiment Properties" (START TIME, END TIME, STATUS CODE, STATUS DETAILS) and a "Summary" section. The bottom status bar indicates "Output saved as 'Dynamometer Clusters'." and includes a "CLOSE" button. The bottom navigation bar contains icons for NEW, RUN HISTORY, SAVE, SAVE AS, DISCARD CHANGES, RUN, SET UP WEB SERVICE, and PUBLISH TO GALLERY.

Microsoft Azure Machine Learning Studio

Oil & Gas - Dynamometer Clustering

In draft

Draft saved at 10:01:11 PM

Search experiment items

Machine Learning

Score

Apply Transformation

Assign Data to Clust...

Score Matchbox Rec...

Score Model

Train

OpenCV Library Modules

Python Language Modules

R Language Modules

Statistical Functions

Dynamometer Card Reading...

K-Means Clustering

Sweep Clustering

Assign Data to Clusters

Experiment Properties

START TIME 9/25/20...

END TIME 9/25/20...

STATUS CODE InDraft

STATUS DETAILS None

Prior Run

Summary

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

Description

Enter the detailed description for your experiment.

Quick Help

Output saved as 'Dynamometer Clusters'.

NEW

RUN HISTORY

SAVE

SAVE AS

DISCARD CHANGES

RUN

SET UP WEB SERVICE

PUBLISH TO GALLERY