**Lab for Chapter 5:**

**Clustering**

Contents

[Exercise 0: Overview **4**](#_bookmark0)

[Exercise 1: Create Model **5**](#_bookmark1)

[Step 1: Get Data 5](#_bookmark2)

[Step 2: Preprocess Data 6](#_bookmark3)

[Exercise 2: Train the Model **7**](#_bookmark4)

[Step 1: Choose and Apply Machine Learning Algorithms 7](#_bookmark5)

[Exercise 3: Test the Model **8**](#_bookmark6)

[Step 1: Predict Over New Data 8](#_bookmark7)

# **Exercise 0: Overview**

Identify what kind of machine learning problem this is.

## “**Group the whole-sale customers into segments**?”

This exercise will create a Segmentation model and classify customers into groups. The simplest process to create any predictive model is

**Create a Model**

1. Get data
2. Preprocess data
3. Define features

**Train the Model**

1. Choose and apply a Machine Learning algorithm

**Test the Model**

1. Predict over new data

*The wholesale customer dataset is available on the UCI Machine Learning Repository. The dataset contains eight columns (referred to as attributes or features) and contains information on the customers of a wholesale distributor, operating in different regions. The columns include annual spending on fresh milk, grocery, frozen, detergents, paper and delicatessen products. In addition, it also includes information on the channel of the customer (hotel/café/restaurant) or retail.*

***Refer to*** [***http://archive.ics.uci.edu/ml/datasets/Wholesale+customers.***](http://archive.ics.uci.edu/ml/datasets/Wholesale%2Bcustomers)

# **Exercise 1: Create a Model**

# **Step 1: Get Data**

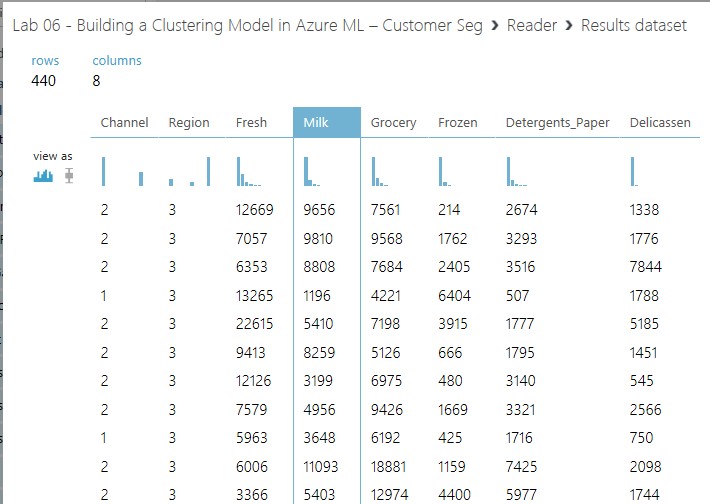
* 1. Create a new experiment and rename it appropriately (Lab 06 - Building a Clustering Model in Azure ML – Customer Seg).
  2. In the **previous lab** exercises we used pre-loaded saved datasets & Azure storage. For this exercise, we will get data from web directly using HTTP source.
  3. Drag a “**Import Data**” module onto the experiment canvas and configure it using below details,

**Data Source**: - Web URL via HTTP

**URL**: - <http://archive.ics.uci.edu/ml/machine-learning->databases/00292/Wholesale%20customers%20data.csv **Data format**: - CSV

**CSV or TSV has header row**: - checked

* 1. Run the experiment.
  2. By clicking the output port of the Import Data module, select ‘Visualize’ to see what we just downloaded from Web and see if there are any issues in the dataset.



# 

# **Step 2: Preprocess Data**

1. We have seen pre-processing in last 2 labs and so will skip this step and assume data is in good shape for us to input to clustering models

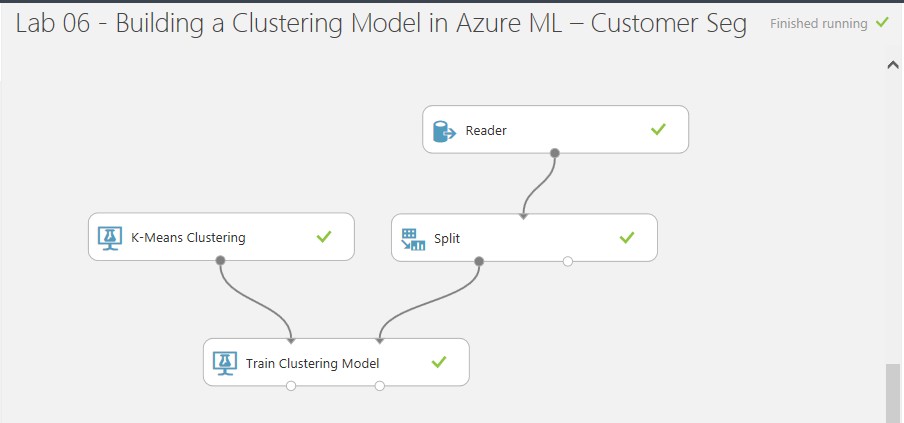
# **Exercise 2: Train the Model**

# **Step 1: Choose and Apply Machine Learning Algorithms**

When constructing a predictive model, you first need to train the model, and then validate that the model is effective. In this experiment, you will build a Clustering model.

In this experiment, you will train a clustering model and use it to classify the customers into groups of similar properties.

1. Split the data into training and testing sets: Select and drag the “**Split Data**” module to the experiment canvas and connect it to the output of the Import Data module.
2. In the properties pane, select ‘Splitting mode’ to ‘Split Rows’. Next, set ‘Fraction of rows in the first output dataset’ to **0.7**. This way, you will use 70% of the data to train the model and hold back 30% for testing. Set Random Seed to 123.
3. Run the experiment. This allows Split module to pass along column definitions to the modules you will be adding next.
4. To select the learning algorithm, expand the Machine Learning category in the module palette to the left of the canvas and then expand ‘Initialize Model’. This displays several categories of modules that can be used to initialize a learning algorithm.
5. For this example, experiment, select the “**K-Means Clustering**” module under the ‘Clustering’ category and drag it to the experiment canvas. For the K-Means Clustering module, configure it to identify four clusters (Number of Centroids = 4), and use metric as ‘Euclidean’ distance measure. Use the ‘Iterations’ number which is set to default as 100.
6. Find and drag the “**Train Clustering Model**” module to the experiment. Click ‘Launch column selector’ and select **All Columns**
7. Connect the output of the “**K-Means Clustering**” module to the left input port of the “Train Clustering Model” module.
8. Also, connect the training data output (i.e. the left port) of the “Split” module to the right input port of the “**Train Clustering Model**” module.
9. Run the experiment.

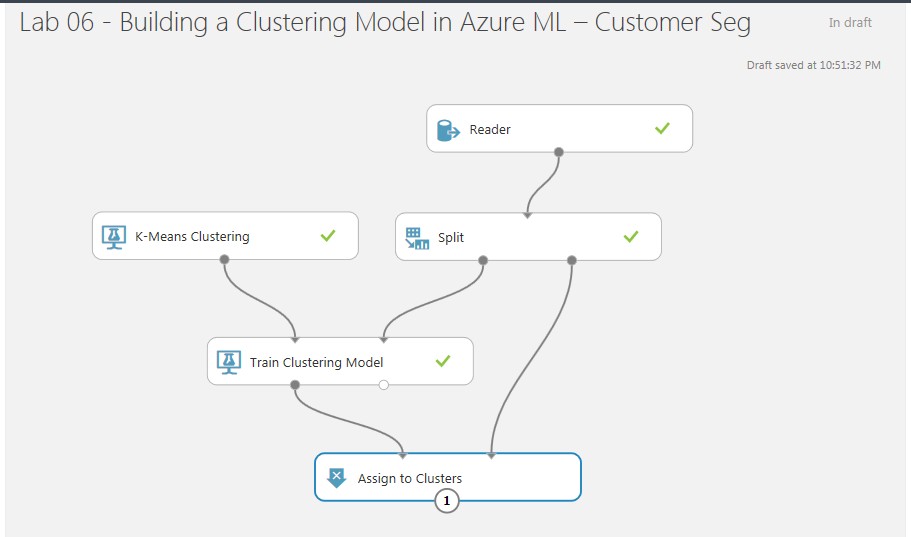


**Exercise 3: Test the Model**

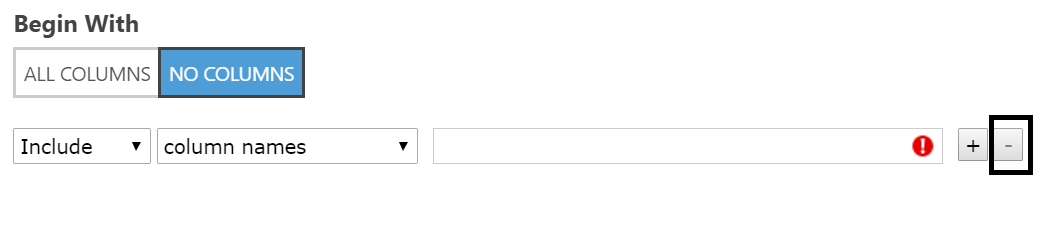
# **Step 1: Predict Over New Data**

Now that you’ve trained the model, you can use it to score the other 30% of your data and see how well your model predicts on unseen data.

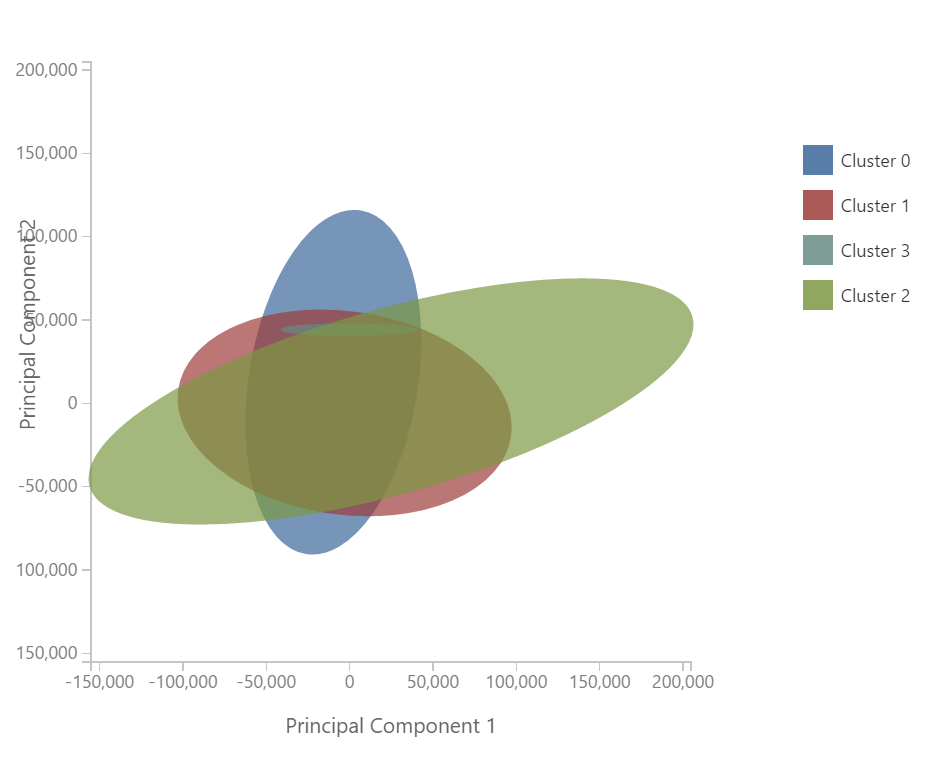
1. Find and drag the “**Assign to Clusters**” module to the experiment canvas. Connect the left output of the “Train Clustering Model” module to the left input port and test data output (right port) of the “Split Data” module to the right input port.



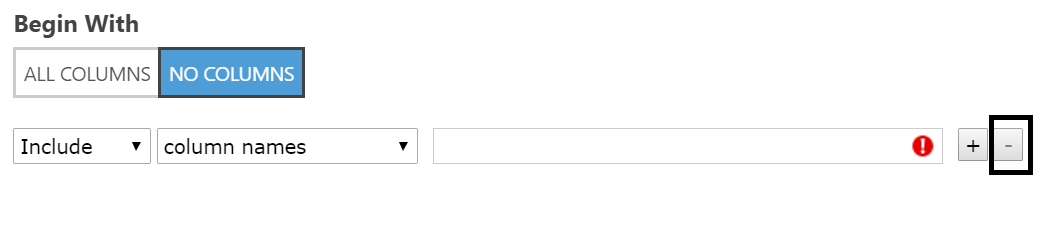
1. Configure the “Assign to Clusters” module as below,
   1. Launch the column selector.
   2. Begin with: “All Columns”
   3. Directly below “Begin With” property, there is a “+” and “-“button. Remove the extra parameter by clicking the minus button as highlighted below. Submit the popup window.



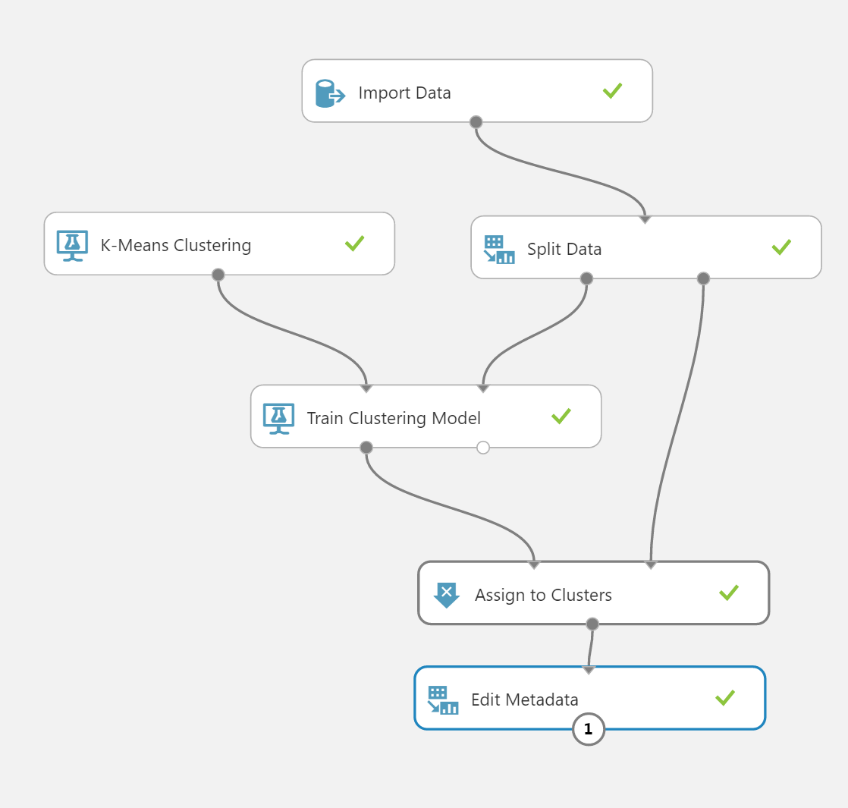
1. Run the experiment and view the output from the “Assign to Clusters” module (by clicking the output port and selecting Visualize). The output will show the clusters graphically.



1. To get the actual data out of it, add an “Edit Metadata“ module and link the output of “Assign to Clusters” module as input to it.
2. Configure the “Edit Metadata” module as below,
   1. Launch the column selector.
   2. Begin with: “All Columns”
   3. Directly below “**Begin With**” property, there is a “+” and “- “button. Remove the extra parameter by clicking the minus button as highlighted below. Submit the popup window.



* 1. Leave “**Data type**”, “**Categorical**” and “**Fields**” to “**Unchanged**”



1. Run the experiment and view the output from the “Edit **Metadata**” module (by clicking the output port and selecting Visualize). The output will be as below.

