Clustering in Azure by Nan(Miya) Wang

Data:

In this lesson we use the dataset ***churn.txt***, containing information on 1,477 customers of a telecommunication firm who has at some time purchased a mobile phone. The customers fall into one of three groups: current customers, involuntary leavers, and voluntary leavers. The file contains information on the customer account including length of time spent on local, long distance and international calls, the type of billing scheme, and a variety of basic demographics, such as age and gender. The data are typical of what is often referred to as a churn example (hence the file name).

The original dataset uses comma-separated txt format. **Machine Learning Studio works better with a comma-separated value (CSV) file, so we'll first convert the dataset into CSV file.**

There are many ways to convert this data. One way is by using the following Windows PowerShell command:

```
cat churn.txt | sc churn.csv
```

If you are using Unix, here are the commands:

```
churn.txt > churn.csv
```

In either case, we have created a comma-separated version of the data in a file named **churn.csv** that we'll use in our experiment.

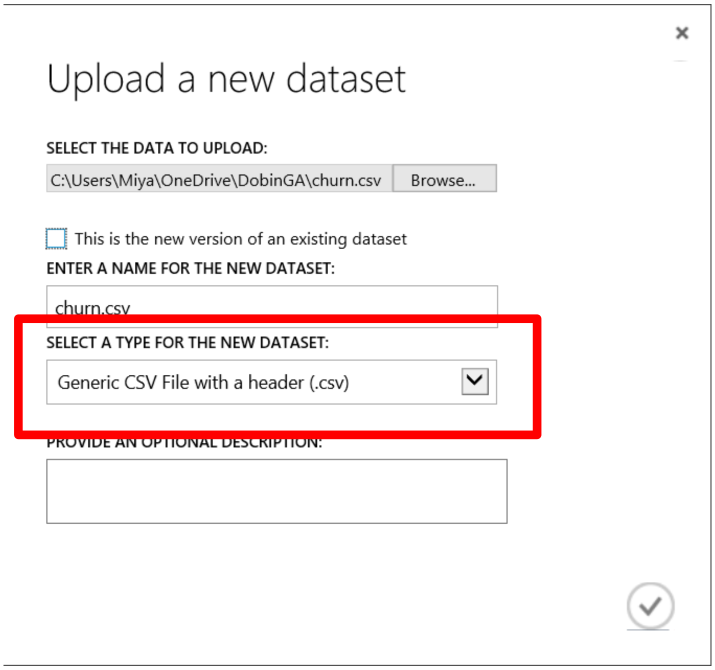
Introduction

We will use the data file *churn.txt*. We will attempt to find natural segments, or clusters of customers to see whether they can be targeted for different promotions and to explore if cluster differences relate to customer status.

Steps:

Prepare the Data

1. Open a browser and browse to http://studio.azureml.net
2. Then sign in using the Microsoft account associated with your Azure ML account (or your school account if available)
3. Create a new dataset named churn.csv by clicking **'NEW'** in the button, then clicking **'DATASET'** and uploading the churn.csv file, as shown in the screenshot below. Click the little check in the circle in the right corner to continue.



Upload a new dataset

SELECT THE DATA TO UPLOAD:

C:\Users\Miya\OneDrive\DobinGA\churn.csv

□ This is the new version of an existing dataset

ENTER A NAME FOR THE NEW DATASET:

churn.csv

SELECT A TYPE FOR THE NEW DATASET:

Generic CSV File with a header (.csv)

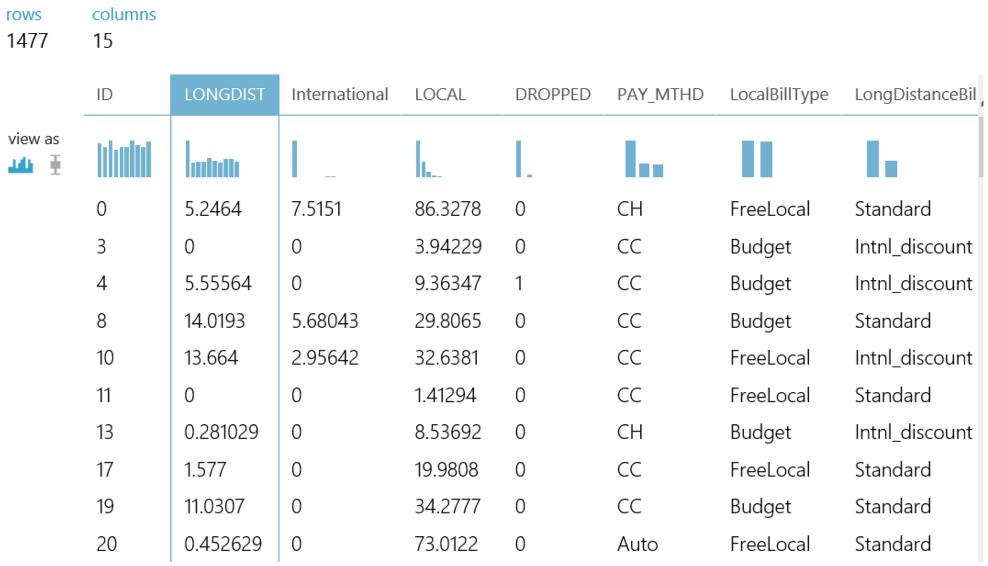
PROVIDE AN OPTIONAL DESCRIPTION:

4. Click '**NEW**' again and create an **EXPERIMENT.** In this stage, just click '**Blank Experiment**'.

A blank experiment canvas would appear.

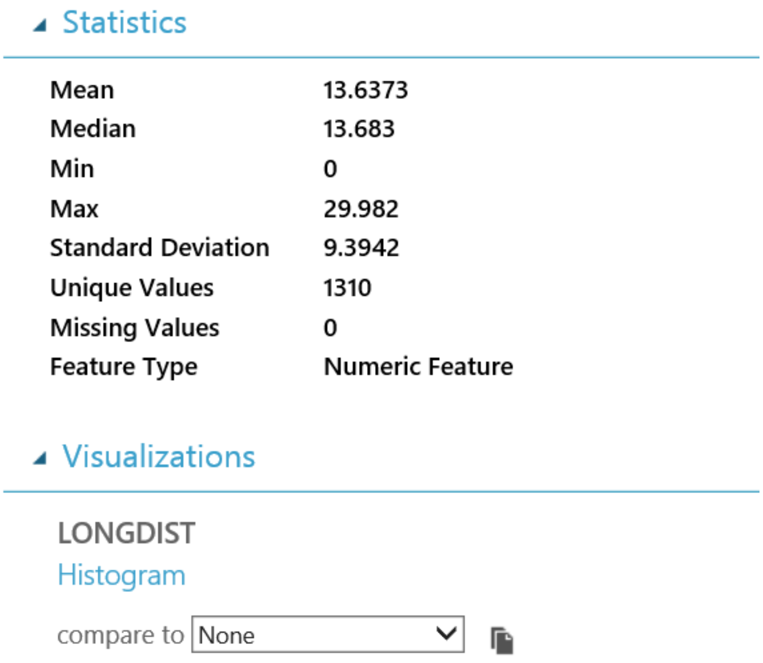
5. Search for 'churrn.csv' in the searching column on the left side and drag it onto the blank canvas.

6. Right-click on the **circle** on churn module and choose **visualize**. Now you can form a general idea of this dataset. Click and go through every data column, statistical facts and visualization of this column automatically formed in the right side.



rows 1477 columns 15

ID	LONGDIST	International	LOCAL	DROPPED	PAY_MTHD	LocalBillType	LongDistanceBil
0	5.2464	7.5151	86.3278	0	CH	FreeLocal	Standard
3	0	0	3.94229	0	CC	Budget	Intnl_discount
4	5.55564	0	9.36347	1	CC	Budget	Intnl_discount
8	14.0193	5.68043	29.8065	0	CC	Budget	Standard
10	13.664	2.95642	32.6381	0	CC	FreeLocal	Intnl_discount
11	0	0	1.41294	0	CC	FreeLocal	Standard
13	0.281029	0	8.53692	0	CH	Budget	Intnl_discount
17	1.577	0	19.9808	0	CC	FreeLocal	Standard
19	11.0307	0	34.2777	0	CC	Budget	Standard
20	0.452629	0	73.0122	0	Auto	FreeLocal	Standard



◢ Statistics

Mean	13.6373
Median	13.683
Min	0
Max	29.982
Standard Deviation	9.3942
Unique Values	1310
Missing Values	0
Feature Type	Numeric Feature

◢ Visualizations

LONGDIST

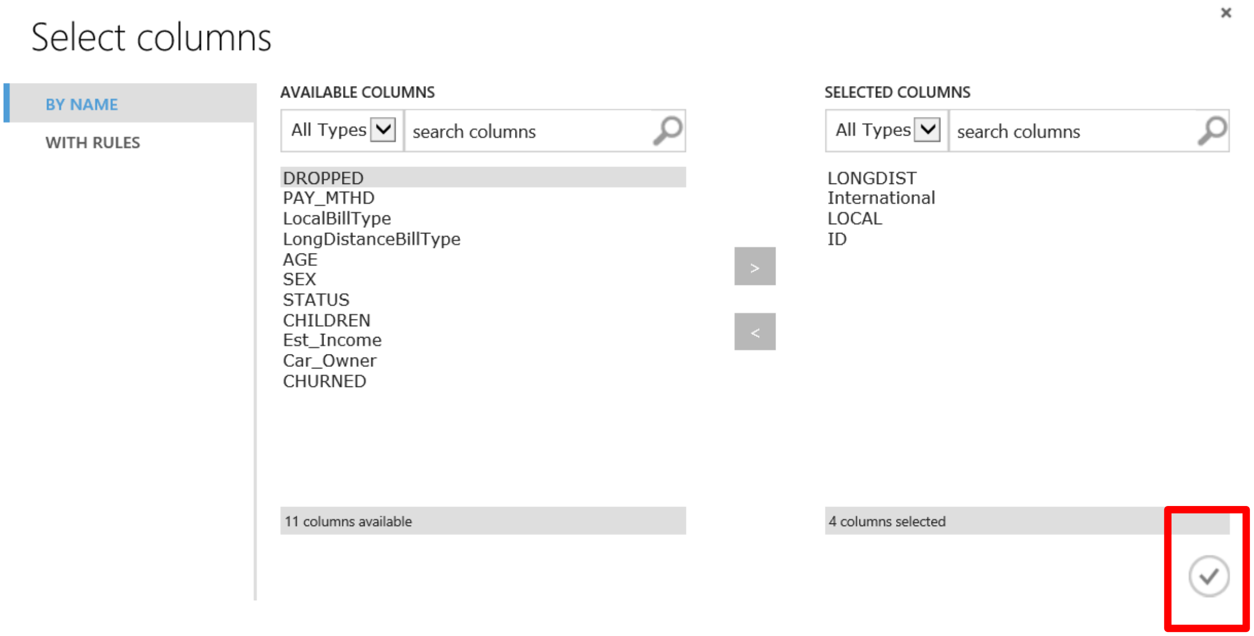
Histogram

compare to None

7. Search for ‘**Project Columns’** and drag it onto the canvas. Connect it to churn dataset by clicking on the circle of the churn module(not letting go of your mouse) and directing the link to the upside circle on the ‘Project Columns’ module.

8. Click on ‘Project Columns’ and then click on **‘Launch column selector’** on the right.

9. Select columns of **LONGDIST, International, LOCAL** corresponding to the amount of time spent on long distance, international, and local telephone calls, in minutes.(As shown below), and column of **ID**. Click on the check circle in the right corner to continue.



Select columns

BY NAME

WITH RULES

AVAILABLE COLUMNS

All Types | search columns

DROPPED
PAY_MTHD
LocalBillType
LongDistanceBillType
AGE
SEX
STATUS
CHILDREN
Est_Income
Car_Owner
CHURNED

11 columns available

SELECTED COLUMNS

All Types | search columns

LONGDIST
International
LOCAL
ID

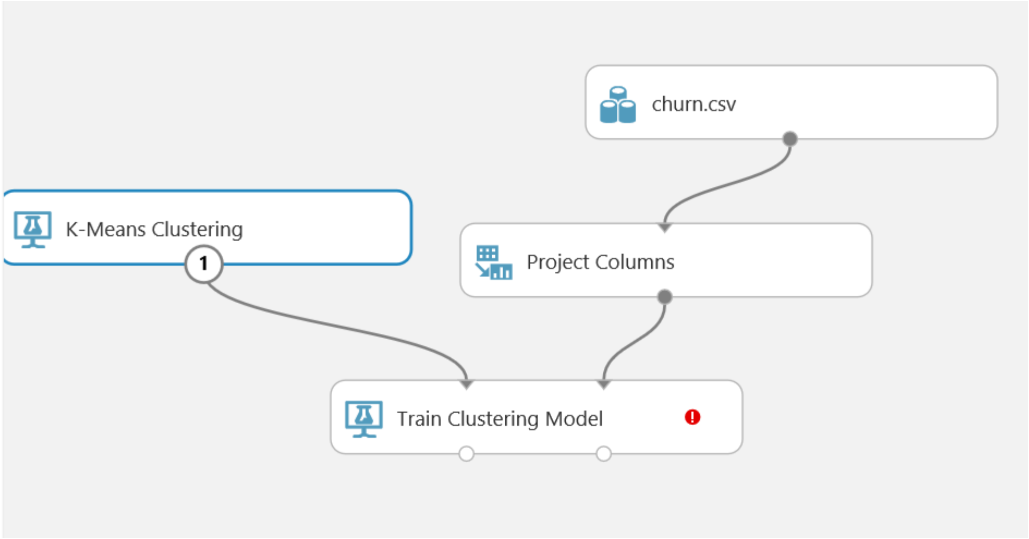
4 columns selected

10. Search for ‘**Train Clustering Model’** and drag it onto the canvas. Connect the right side circle of this module to bottom circle of ‘**Project Columns’** module.

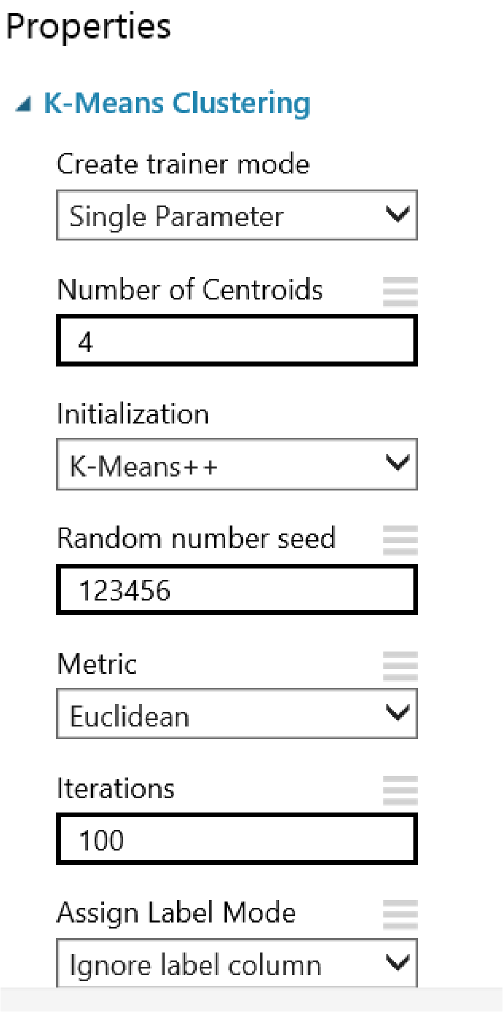
11. Click on the ‘Train Clustering Model’ and then click on ‘Launch Column Selector’ on the right. Choose all the columns **except ID.** Then continue by clicking the check in the right corner.

12. Search for ‘**clustering’** and drag **‘K-Means Clustering’** onto the canvas. (Azure offers only one clustering algorithm by default, which is K-means.)

13. Connect the only circle on the ‘K-Means’ module to left-up side of ‘Train Clustering Model’ module. Now your experiment should look like this:



14. Click on the K-Means clustering module to tune parameters. You can set customized parameters in the right column. Here is an example.



Properties

K-Means Clustering

Create trainer mode: Single Parameter

Number of Centroids: 4

Initialization: K-Means++

Random number seed: 123456

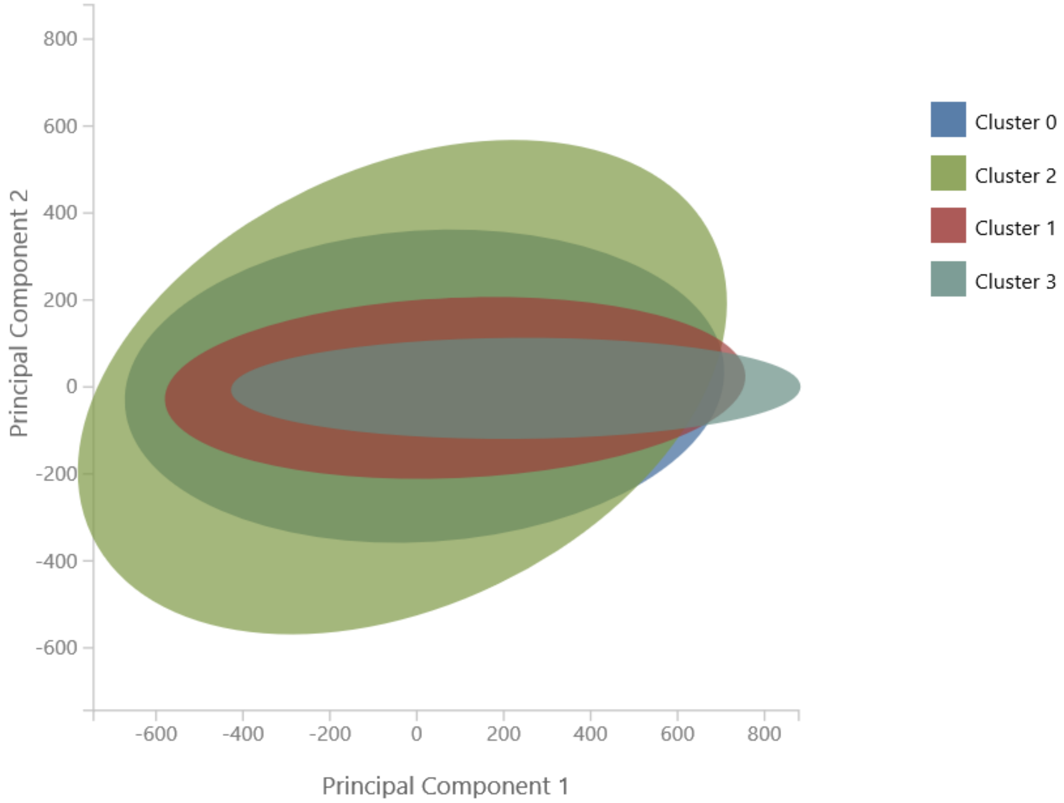
Metric: Euclidean

Iterations: 100

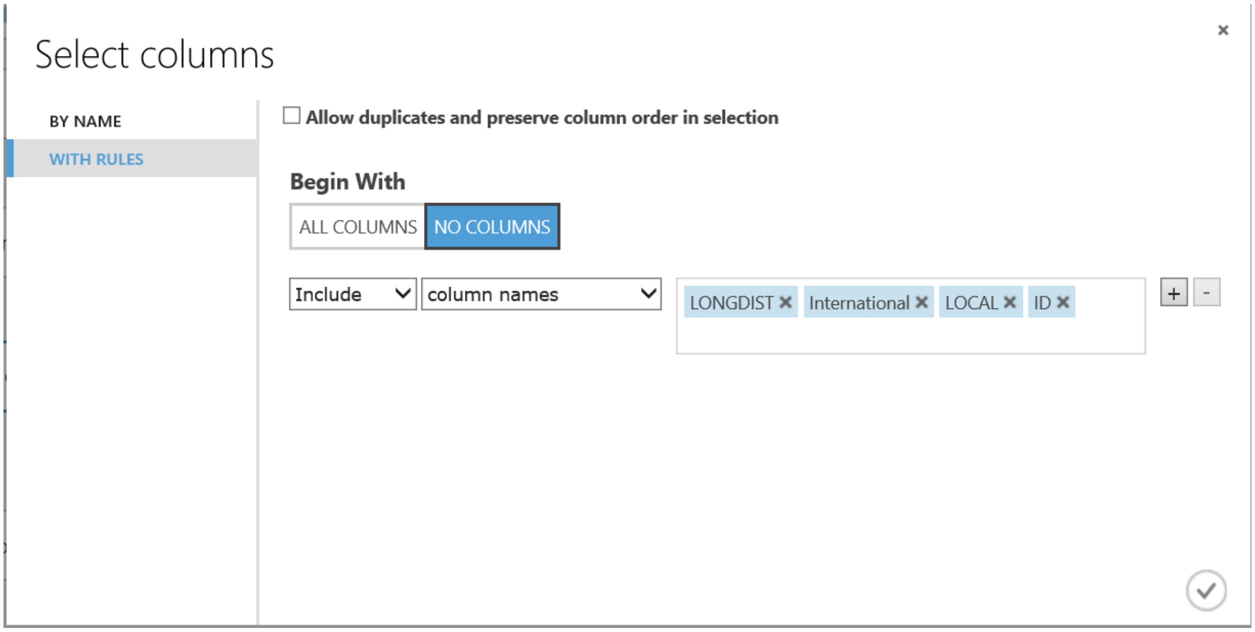
Assign Label Mode: Ignore label column

15. Click on the ‘Run’ button to train the model.

16. When finished running, right click on the right circle of the module “Train Clustering Module” and select ‘Visualize’. You can see the picture below, which uses PCA (principal component analysis) technique to present clusters,



17. Search for ‘**Assign to Clusters’** module and drag it onto the canvas. Connect it to the module ‘Training Cluster Model’ and select all the columns by clicking on ‘Launch column selector’ on the right.



Select columns

BY NAME

WITH RULES

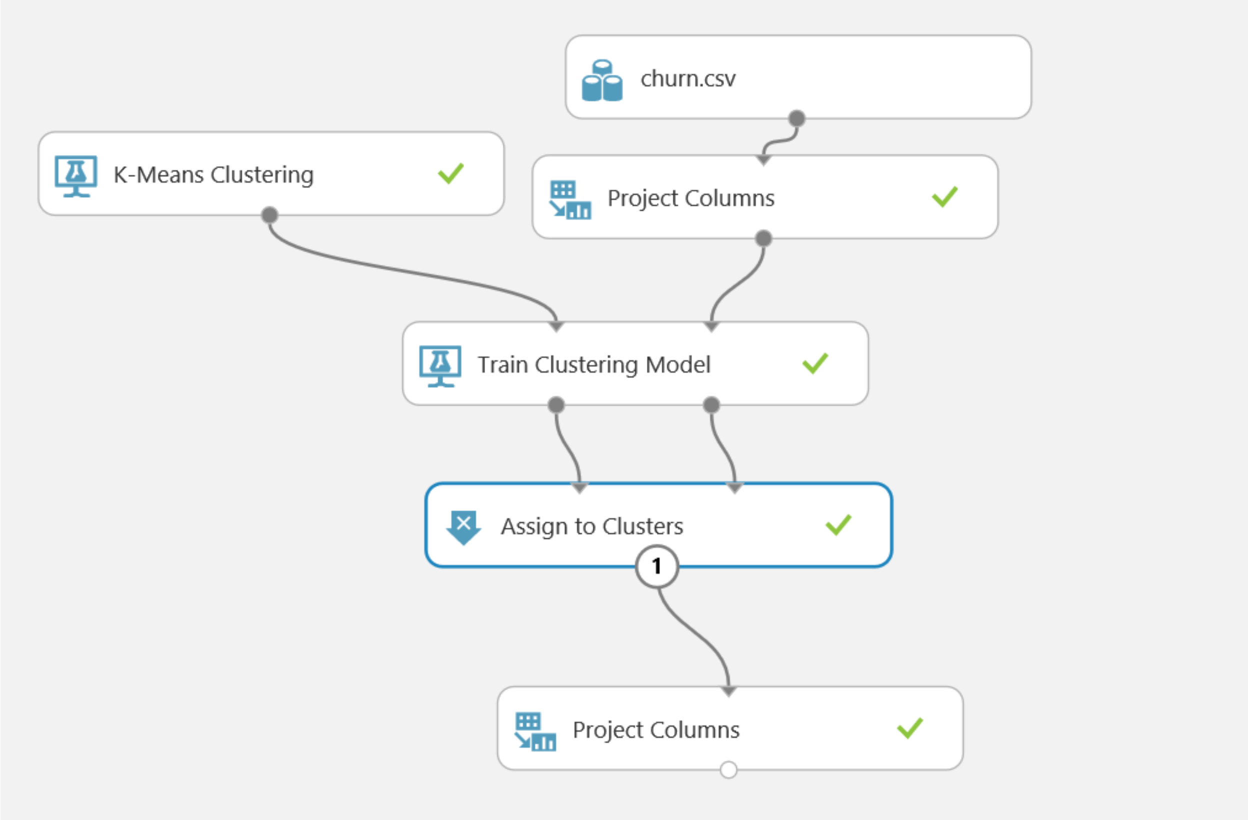
□ Allow duplicates and preserve column order in selection

Begin With

Include | column names | LONGDIST × International × LOCAL × ID ×

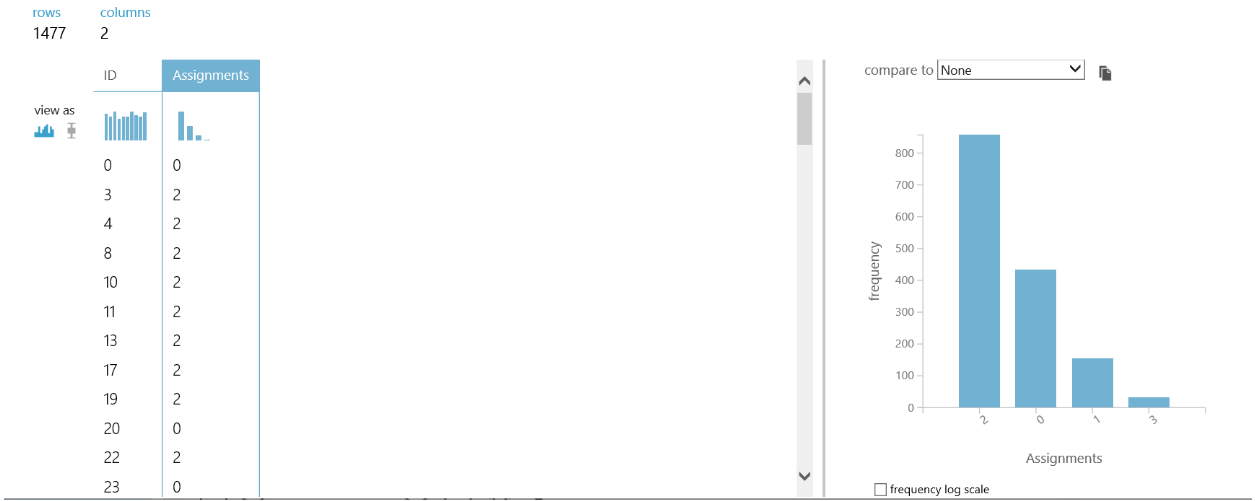
18. Run the experiment again.

19. When finished, add ‘Project columns’ again to the experiment canvas and connect it to the ‘Assign to clusters’. Now your experiment should look like this.



20. Select ID and Assignments columns for the ‘Project Columns’ module and visualize it.

Now you can see customers belong to which clusters.



rows 1477

columns 2

ID	Assignments
0	0
3	2
4	2
8	2
10	2
11	2
13	2
17	2
19	2
20	0
22	2
23	0

compare to None

21. Notes: visualization in azure in quite limited. But you can use customized python codes or R codes to do visualization and thus results evaluation. Explore it by yourself!