Exercise 4: Training a Clustering Model

So far you have trained two supervised machine learning models: one for regression, and one for classification. Clustering is an example of *unsupervised* learning; in other words, training a predictive model with no known labels. In this exercise, you will copy an existing training experiment from the Azure AI Gallery and run it to train a K-Means clustering model that segments the lemonade sales customers into clusters based on similarities in their features.

Copy an Experiment from the Gallery to your Workspace

6. Navigate to https://gallery.cortanaintelligence.com/Experiment/Lemonade-Clustering-Customers. This opens the **Lemonade - Clustering Customers** example in the Azure AI Gallery

7. Click **Open in Studio** to copy the experiment and its dataset to your Azure Machine Learning Studio workspace. If you are prompted, sign in using your Microsoft account, and then select the region and workspace for your Azure Machine Learning Studio subscription.

8. When the experiment opens in Azure Machine Learning Studio, rearrange the panes and zoom level so you can see the modules it contains. The experiment data flow starts with a dataset named **Customers.csv**.

* 1. 9. Visualize the output of the **Customers.csv** dataset, and note that it contains observations for 109 customers, including the following features:
  2. • **CustomerID**: A unique identifier for each customer.
  3. • **Name**: The customer’s full name.
  4. • **Age**: The age of the customer.
  5. • **AvgWeeklySales**. The average number of sales to this customer per week.
  6. • **AvgDrinks**: The average number of drinks purchased by this customer per sale.

1. 10. Review the rest of the experiment, noting that unlike the supervised learning experiments you have conducted previously, there is no step to split the data and withhold a set for testing. This is because in an unsupervised learning model, there is no known label with which to test the predictions.
2. 11. Select the **K-Means Clustering** module and view its settings in the **Properties** pane.

Note that the K-Means clustering algorithm is configured to initialize 3 random centroids, and then perform 200 iterations of assigning observations to their nearest centroid and then moving the centroid to the middle of its cluster of observations.

1. Run the Experiment and View the Results
2. 1. Run the **Lemonade – Clustering Customers** experiment and wait for it to complete.
3. 2. When the experiment has finished running, visualize the **Results dataset** (right) output of the **Train Clustering Model** module to view the Principle Component Analysis (PCA) visualization for the results.
4. The PCA ellipses for the clusters are all oriented in different directions, indicating that there is some separation between them – though clusters 0 and 1 are not as well separated as clusters 1 and 2.
   1. 3. Visualize the output of the **Apply SQL Transformation** module at the end of the experiment, and note that it contains the following new fields:
   2. • **Assignments**: The cluster to which this observation has been assigned (0, 1, or 2).
   3. • **DistancesToClusterCenterno.0**: The distance from this observation to the center of cluster 0.
   4. • **DistancesToClusterCenterno.1**: The distance from this observation to the center of cluster 1.
   5. • **DistancesToClusterCenterno.2**: The distance from this observation to the center of cluster 2.
5. Each observation has been assigned to the cluster to which it is closest.

*Challenge*

1. Note the **Assignments** value indicating the cluster to which customer 1 (Marva Cardenas) is assigned.