**KMeans Clustering**

KMeans is an unsupervised clustering technique used for segmentation or understanding the underlying data pattern as part of exploratory data analysis.

In unsupervised learning, we have only the features X and no labels y available in our dataset, hence by the similarity of the data from each other, clusters are formed. The data points in each cluster will be similar to each other and the data points between 2 different clusters will be low on similarity. This is the cost function of the KMeans clustering known as the within cluster sum of squares. The similarity measure is based on the Euclidean distance calculated between the datapoints.

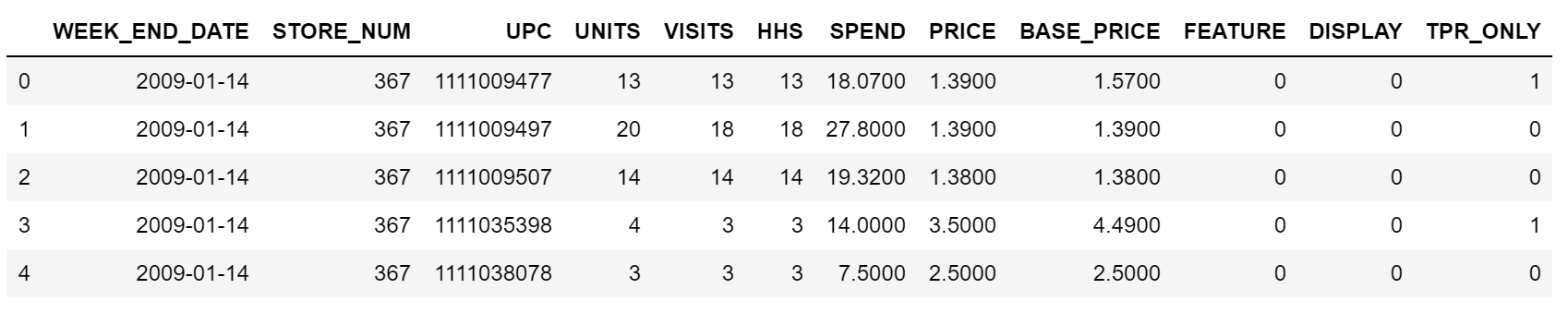
KMeans look for k which is the number of centroids. A centroid represents the center of the cluster.

**KMeans working with Dunhumby Breakfast at the Frat dataset using Python**

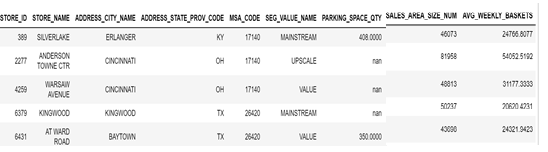
The Dunhumby Breakfast at the Frat dataset contains sales data over 156 weeks for the products mouthwash, pretzels, frozen pizza, and boxed cereal collected over 77 stores. We will use the clustering technique to group the stores based on certain features.

Step1: As the first step we will import all the libraries needed for visualization and reading data and performing the KMeans Clustering

Step2: Reading the Transaction/Sales data. This consists of weekly sales information for each store and product. We have UNITS, VISITS, HHS(purchasing households),SPEND,PRICE and BASE\_PRICE

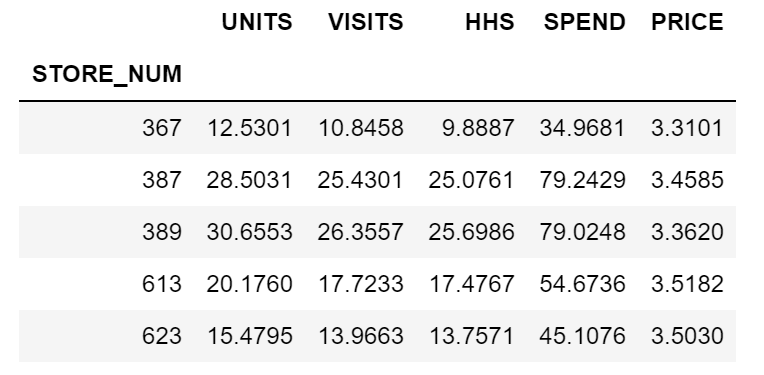


Step2: Reading the Store dataset.We have additional information of SALES\_AREA\_SIZE\_NUM(square footage of store) and AVG\_WEEKLY\_BASKETS(average weekly baskets sold in the store)

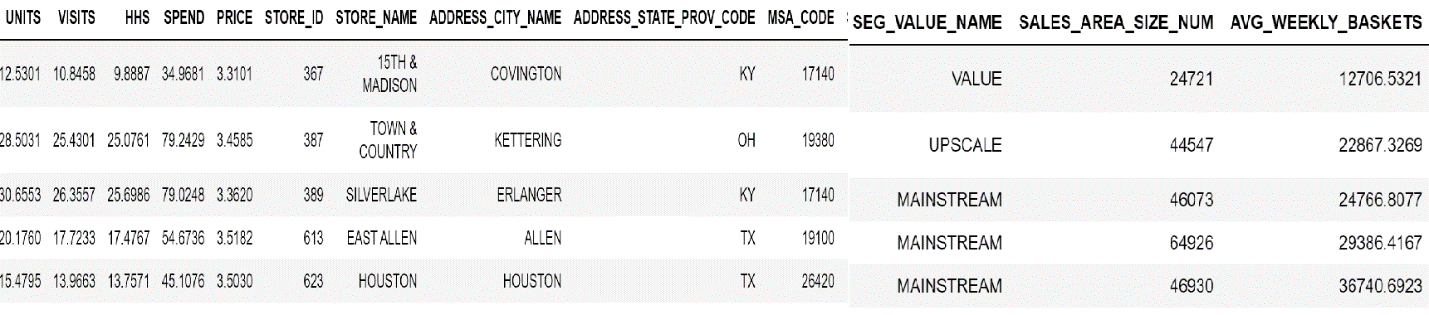


Step 3: Checking for null values and duplicates. For the Store data, 2 stores had duplicate entry and PARKING\_SPACE\_QTY has lot of null values.

Step4: From the Dunhumby transaction data, we will create a new dataframe with the average values of the features UNITS, VISITS, HHS,SPEND and PRICE by the STORE\_NUM



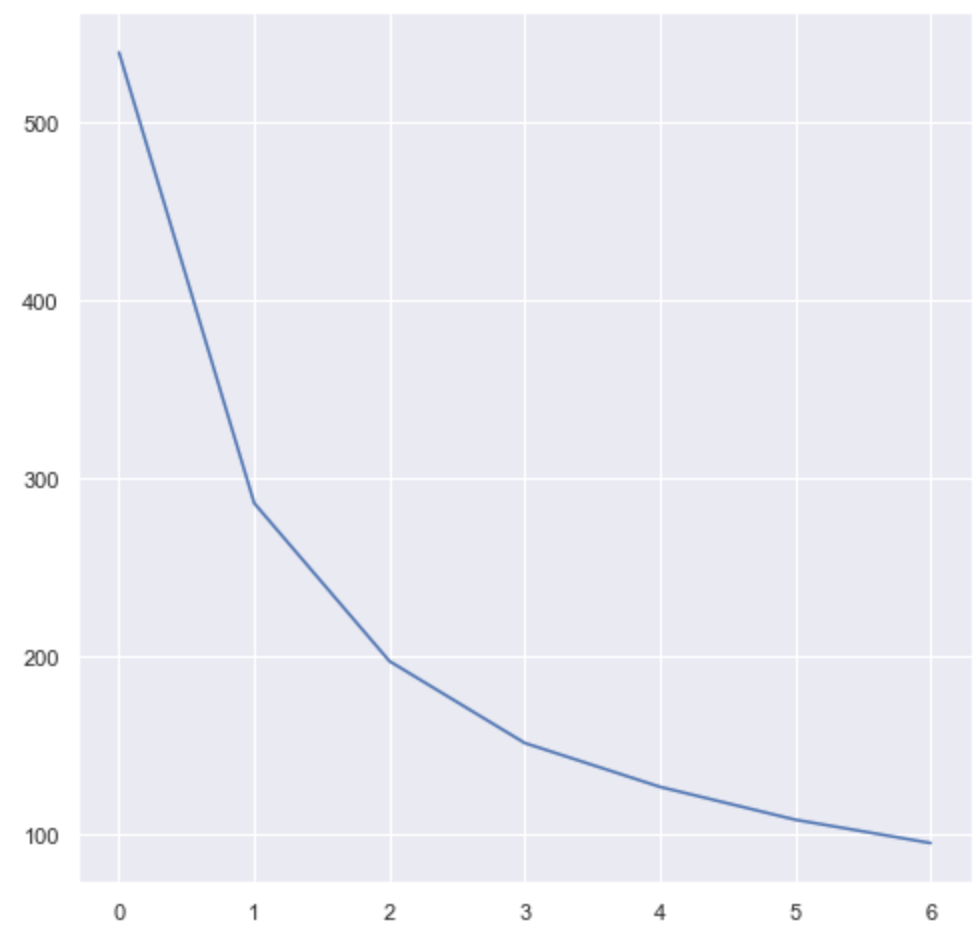
Step5: We will combine the aggregated data with the store data to make the final dataset



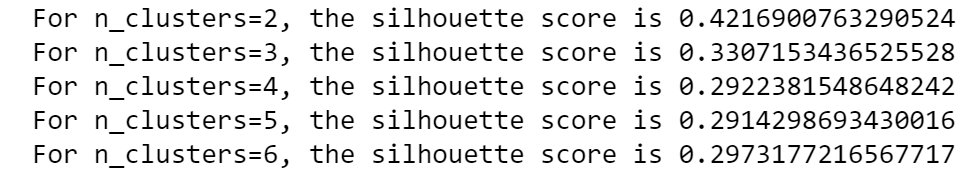
Step6: The following features ['UNITS','VISITS','HHS','SPEND','PRICE','SALES\_AREA\_SIZE\_NUM','AVG\_WEEKLY\_BASKETS'] will be selected for performing the store segmentation

Step7: Since Euclidean distance is used to measure the similarity/dissimilarity, we got to make sure all the features are within the same range. Hence rescaling all the columns so that the values lie between 0 and 1 using the standardization scaling

Step 8: Using the Elbow curve method to check for the ideal number of clusters. We will analysis our grouping of stores with 2, 3 and 4.



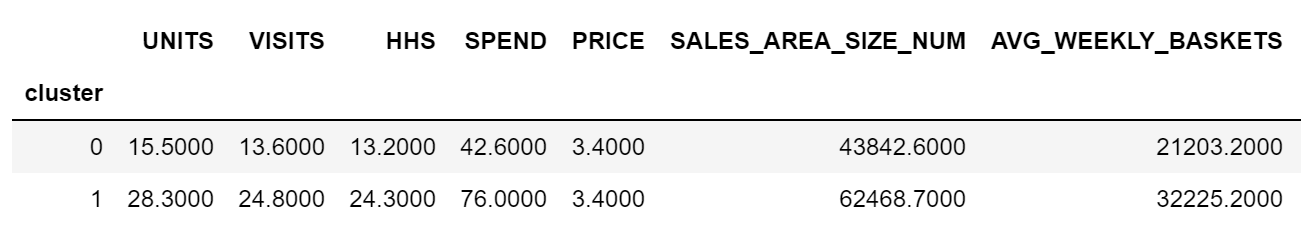
Step 9: Let’s check the Silhouette score for the different clusters. The Silhouette score lies between -1 and 1. For good separation of the data points between different clusters, the score should tend more towards 1. Here we can see the separation for 2 clusters is good with value 0.42 and then with 3 clusters with value of 0.33.



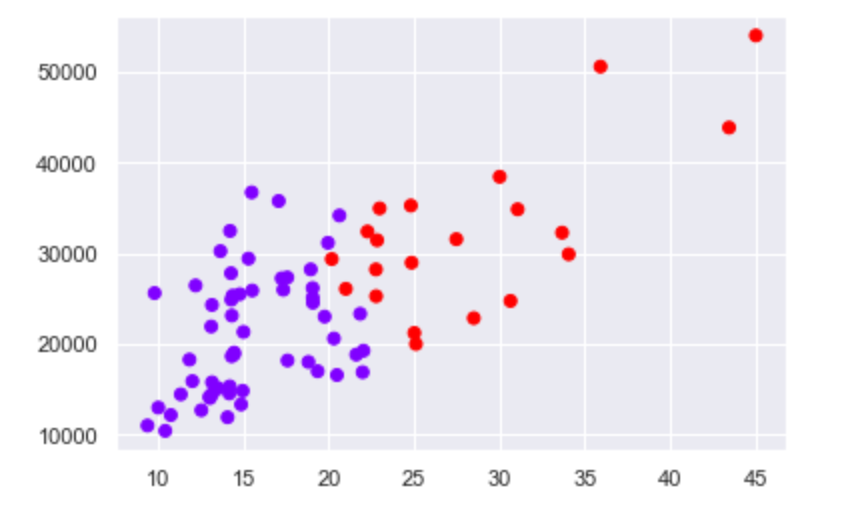
Step 11: Based on the Elbow curve and the Silhouette score, let’s explore the Stores segmentation with 2, 3 and 4 clusters.

Cluster 2 Analysis…

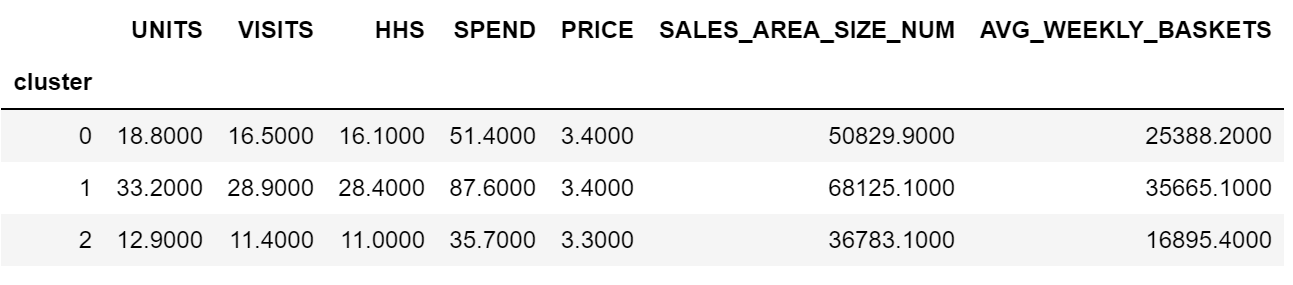
Here the centroid datapoints of each cluster as shown below: The first cluster is low on UNITS, VISITS, HHS, SPEND, SALES\_AREA\_SIZE\_NUM and AVG\_WEEKLY\_BASKETS compared with the second cluster.



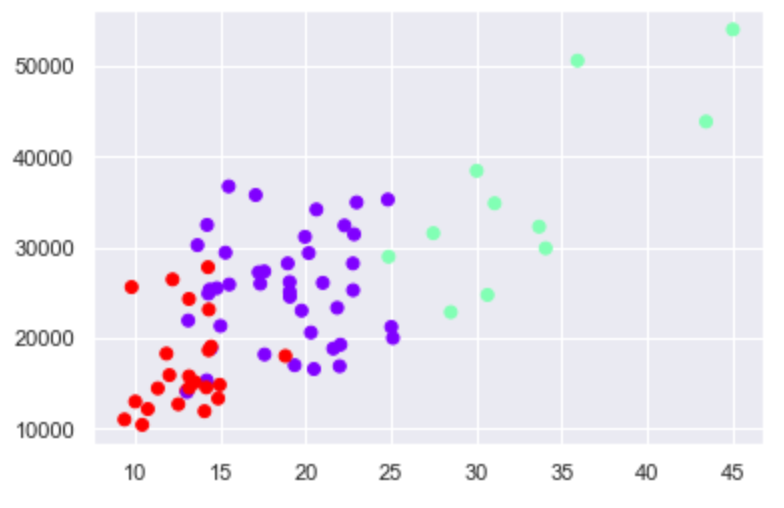
The clusters visualization with 2 features AVG\_WEEKLY\_BASKETS on the y axis and UNITS on the x axis:



Cluster 3 Analysis…

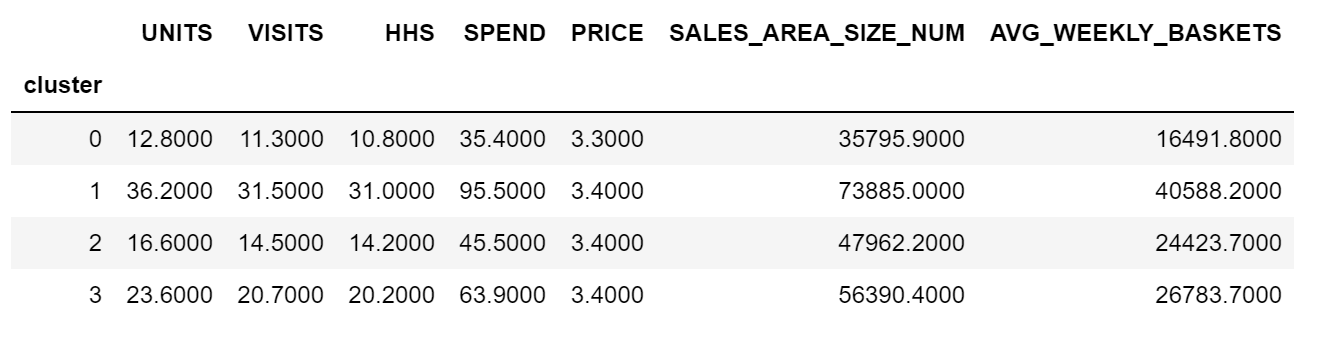
Here the centroid datapoints of each cluster as shown below: Here we can clearly see that the third cluster is low on UNITS, VISITS, HHS, SPEND, SALES\_AREA\_SIZE\_NUM, AVG\_WEEKLY\_BASKETS. Next is the first cluster. The second cluster is highest across all these features. 

The clusters visualization with 2 features AVG\_WEEKLY\_BASKETS on the y axis and UNITS on the x axis:

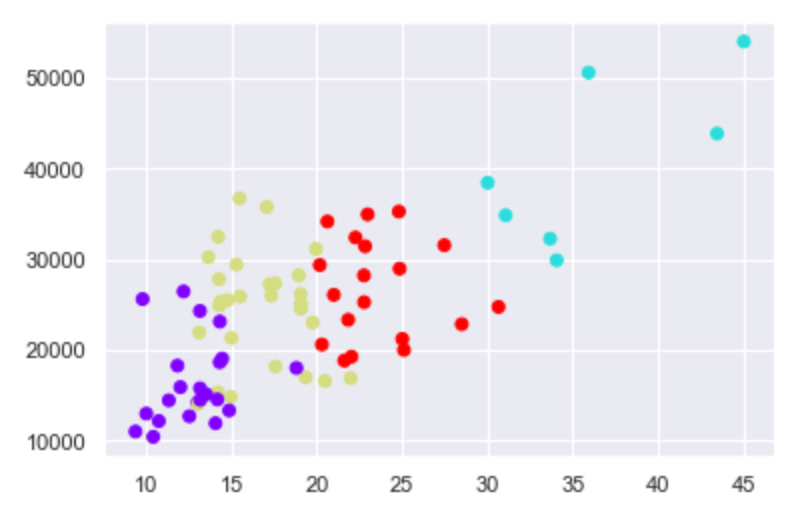


Cluster Analysis with 4 clusters:

Here the centroid datapoints of each cluster as shown below: Cluster[0] is low on UNITS, VISITS, HHS, SPEND, SALES\_AREA\_SIZE\_NUM, AVG\_WEEKLY\_BASKETS. Cluster[2] is higher compared to cluster[0] on the same set of features. Cluster[3] is higher then cluster[2]. Cluster[4] is the highest selling stores.



The clusters visualization with 2 features AVG\_WEEKLY\_BASKETS on the y axis and UNITS on the x axis:



Finally we can either go with 3 or 4 clusters based on our actual business need. As a next step we can also build a predictive model, where a new store coming can be predicted to belong to one of the cluster based on the given features.