# Objective

The purpose of the assignment is to learn about concepts related to unsupervised learning techniques, specifically kMeans Clustering. The data to apply clustering involves the historical transactions of customers for a shopping mart. The assignment also provides valuable hands-on experience with the customer segmentation.

# Description of Raw Data

For this assignment 2 separate datasets made available which represents 2 different branches of the shopping mart. Dataset named “sales211” representing the branch with code 211 is used for further analysis. Above mentioned dataset consisted of 1,726,083 rows and 17 columns. The meta data for the features are as follows:

**TRANSACTION\_RK:** Unique transaction id for each item purchased

**CALENDAR\_DT:** Date of transaction formatted as “dMthyyyy”

**DATE:** Date of transaction formatted as “yyyy-mm-dd”

**TIME:** Time of transaction formatted as “hh:mm:ss”

**TRANSACTION\_TM:** Time of transaction formatted as “hh:mm:ss”

**ITEM\_SK:** Unique id for a given product

**RETAIL\_OUTLET\_LOCATION\_SK:** Branch code of given location

**POS\_TERMINAL\_NO:** Pos Payment Related Information

**CASHIER\_NO:** Unique code of the cashier handled the transaction

**ITEM\_QTY:** Quantity of items purchased

**ITEM\_WEIGHT:** Weight of items purchased

**SALES\_UOM\_CD:** Unit of measure for the product

**SELLING\_RETAIL\_AMT:** Cost of given product/s purchased

**PROMO\_SALES\_IND\_CD:** Promotion code for the product

**STAPLE\_ITEM\_FLG:** Whether the product staple flagged

**REGION\_CD:** Region code of the store

**CUSTOMER\_SK:** Unique customer id

# Exporting Data

The raw transaction data obtained from the SMU Server is queried for unique customers and unique products. By doing so, 2 separate tables are created for top 2000 customers and products generating the maximum revenue. This limitation is made due to performance considerations while clustering the subsets of transaction data. Scaling the analysis for the overall data is crucial for creating real life business value. However, since the scope of this assignment is to develop an intuition for unsupervised learning, using a subset is preferred. The queries for [product](#_PRODUCT_QUERY) and customer tables for clustering analysis can be found in the appendix section of the report.

# Loading Data

# Appendix

## PRODUCT QUERY

USE ca\_irfanoglu;

CREATE TABLE productCluster AS

SELECT ITEM\_SK AS ITEM\_ID,

SUM(SELLING\_RETAIL\_AMT) AS TOTAL\_REVENUE,

COUNT(ITEM\_SK) AS VISITS\_PURCHASED,

COUNT(DISTINCT( CUSTOMER\_SK)) AS DISTINCT\_CUSTOMERS,

AVG( SELLING\_RETAIL\_AMT / ITEM\_QTY) AS AVG\_PRICE

FROM dataset01.sales211

WHERE CUSTOMER\_SK > 1 AND ITEM\_QTY > 0 AND SELLING\_RETAIL\_AMT > 0

GROUP BY ITEM\_SK

ORDER BY TOTAL\_REVENUE DESC

LIMIT 2000;

## CUSTOMER QUERY

USE ca\_irfanoglu;

CREATE TABLE customerCluster AS

SELECT CUSTOMER\_SK,

SUM(SELLING\_RETAIL\_AMT) AS TOTAL\_REVENUE,

SUM(ITEM\_QTY) AS TOTAL\_ITEMS,

COUNT(DISTINCT(ITEM\_SK)) AS DISTINCT\_ITEMS,

COUNT(DISTINCT(TRANSACTION\_RK)) AS TIMES\_VISITED,

MAX(DATE) as MOST\_RECENT\_VISIT,

SUM(SELLING\_RETAIL\_AMT) / NULLIF(COUNT(DISTINCT(TRANSACTION\_RK)),0) AS AVG\_SPENT\_PER\_VISIT

FROM dataset01.sales211

WHERE CUSTOMER\_SK > 1 AND ITEM\_QTY > 0 AND SELLING\_RETAIL\_AMT > 0

GROUP BY CUSTOMER\_SK

ORDER BY TOTAL\_REVENUE DESC

LIMIT 2000;