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 living on 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. The queries for product and customer tables for clustering analysis can be found in the appendix section of the report

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;
```