Help

Selected Columns:

Feature Name	Measurement	Description
ITEM_SK	N/A	Unique field for identifying items
DISTINCT_CUSTOMERS	COUNT DISTINCT	Total number of customers who have purchased the item at some point.
TOTAL_REVENUE	SUM	Total revenue for that item
BASKETS	COUNT DISTINCT	Total number of baskets the product was found in. For the purposes of this analysis, visits, baskets, and transaction are used synonymously. For instance, a person can have multiple visits to the same store on
AVERAGE_PRICE	AVG	The average price of the item. This compensates for changes in price across the dataset and does not include instances where the price is $0.00, which may indicate a special price calculated for BOGO or coupons.

library(ggplot2) → It is used to make millions of plots.

library(GGally) → It extends ggplot2 by adding several functions to reduce complexity.

library(DMwR) → This includes functions and data accompanying the book “Data Mining with R, learning with case studies”

set.seed(5580) →set.seed(seed) Set the seed of R’s random number generator, which is useful for creating simultions or random objects that can be reproduced.

prod <- read.csv("~/Downloads/productcluseter.csv") → This is used to read the csv file. You can use option "import Dataset" to import CSV file if you are getting path error.

View(prod) → To View the loaded dataset

ggpairs(prod[, which(names(prod) != "ITEM_SK")], upper = list(continuous = ggally_points), lower = list(continuous = "points"), title = "Products before outlier removal") → To Visualize data

Syntax:

ggpairs(data[rows, columns], upper=list(continuos = “ggally_points”), lower=list(continuous=”points”, title =”Any text”)

Here we are not eliminating any rows, so we are not mentioning anything to consider all rows.

From column we don't want to keep ITEM_SK so we are mentioning the condition which(names(prod) != "ITEM_SK").

upper and **lower** are lists that may contain the variables 'continuous', 'combo', 'discrete', and 'na'.

continuous -> exactly one of ('points', 'smooth', 'smooth_loess', 'density', 'cor', 'blank'). This option is used for continuous X and Y data.

Title is used to give any heading to the plot.

More details:
https://www.rdocumentation.org/packages/GGally/versions/1.4.0/topics/ggpairs
boxplot(prod$BASKETS) → For Box and Whisker plot. here prod is dataset and BASKETS is column

prod.clean <- prod[prod$ITEM_SK != 11740941,] →

Now we want to remove only row with ITEM_SK = 11740941 from the prod. So we are not mentioning anything after comma, it means it will keep all columns of prod.

ggpairs(prod.clean[,which(names(prod.clean)!="ITEM_SK")], upper = list(continuous = ggally_points), lower = list(continuous = "points"), title = "Products after removing ITEM_SK= 11740941 (Bananas)") → Visualize after removing outliers

prod.scale = scale(prod.clean[-1]) → Normalize data using scale and exclude ITEM_SK column. -1 will remove first column that is ITEM_SK and keep all other.

```
withinSSrange <- function(data,low,high,maxIter)
{
  withinss = array(0, dim=c(high-low+1));
  for(i in low:high)
  {
    withinss[i-low+1] <- kmeans(data, i, maxIter)$tot.withinss
  }
  withinss
}
```

→ Define withinSSrange

plot(withinSSrange(prod.scale,1,50,150)) → Elbow plot to determine the optimal number of clusters between 1 and 50. Here 150 is number of iterations.

pkm = kmeans(prod.scale, 5, 150) → K-means using k=5 for products based on results of elbow plot. Here 150 is number of iterations.

prod.realCenters = unscale(pkm$centers, prod.scale) → Denormalize data by reversing scale function

clusteredProd = cbind(prod.clean, pkm$cluster) → Bind clusers to cleansed Data

plot(clusteredProd[,2:5], col=pkm$cluster) → Visualizing clusering results. Here we want all rows so we are not mentioning anything but we want columns only from 2 to 5 (we don't want to visualize first column - ITEM_SK).

write.csv(clusteredProd, file = "/Users/trishlashah/Downloads/results.csv", col.names = FALSE) → Export result

Analysis example based on the result.

Customer Segment	Description	Recommendation
Cluster 4 – Champions	• High frequency buyers • Bought recently • High total and average spend • Purchase a wide array of products	Customers are already highly engaged. Continue to monitor.
Cluster 1 – Require Attention	• High frequency buyers • Have not bought recently • High total and average spend • Purchase a wide array of products	Focus on getting these customers back into the store through targeted marketing and promotions.

Cluster 5 – High Potentials	• Low frequency buyers • Bought recently • Medium total and average spend • Purchase a wide array of products	Focus on engaging the customer, upselling, and increasing the amount they're spending.
Cluster 2 – Bargain Hunters	• Medium frequency buyers • Bought recently • Low total and average spend • Purchase few products	Targeted promotions on higher value products would help to increase the spend and frequency of these buyers.
Cluster 3 – Impulse Buyers	• Medium frequency buyers • Bought semi-recently • Low total spend • High average spend • Purchase few products	Offer complimentary products and promotions to increase total spend.