---
title: "Assignment5_AD699"
author: "Assel_Kassenova"
date: "4/28/2023"
output: pdf_document
---

1. Describe “Groceries”
```{r}
library(arules)
library(Matrix)
data('Groceries')

class(Groceries)
dim(Groceries)
```

2. Frequency barplot. The class of "Groceries" is a transaction dataset. The "Groceries" dataset contains 9835 rows (transactions) and 169 columns (items).

```{r}
library(tibble)
library(dplyr)
item_freq <- itemFrequency(Groceries)

# Get top 12 items
top_12_items <- head(sort(item_freq, decreasing = TRUE), 12)
top_12 <- as.data.frame(top_12_items)

top  <- rownames_to_column(top_12, var = "index")

#barplot of the top 12 items
barplot(top$top_12_items, names.arg = top$index, col = "blue", las = 2,
        main = "Top 12 Most Common Grocery Items",
         ylab = "Item Frequency")
```


3. Now, create a subset of rules that contain any grocery item of your interest

```{r }
grocery_rules <- apriori(Groceries, parameter = list(confidence = 0.5, minlen = 2))

butter_rules <- subset(grocery_rules, lhs %in% "butter")

inspect(butter_rules)
summary(butter_rules)

milk_rules <- subset(grocery_rules, rhs %in% "whole milk")
# Print the sugar rules and their support, confidence, and lift
inspect(milk_rules)
```

4. In a sentence or two, explain what meaning these rules might have for a store like Star Market. What could it do with this information?

Answer: First of all i set the  confidence parameter to 0.5, meaning that i'm have a confidence of at least 50%.
For example on butter left rules output is:  {butter, domestic eggs} => {whole milk}      support  0.005998983  confidence 0.6210526 , which means butter is bought along with whole milk 0.5 % time with confidence of 60%. It is a great insight that can help Star Market to organize their products shelf accordingly , set the bundle price promotions.

```{r}
library(arulesViz)

rules <- apriori(Groceries, parameter = list(support = 0.001, confidence = 0.8, minlen = 2))
subset_rules <- subset(rules, lhs %in% "sugar" | rhs %in% "whole milk")
three_rules <- sample(subset_rules, 3)

plot(three_rules, method = "scatterplot")
```

```{r}
library(arulesViz)
data("Groceries")

# Create a subset of rules containing the grocery item of interest
pastry_rules <- apriori(Groceries, parameter = list(supp = 0.001, conf = 0.2, target = "rules"), appearance = list(lhs = c("pastry"), default = "rhs"))

# Generate a scatter plot of three rules involving the grocery item
plot(pastry_rules, method = "scatter", jitter = 0.2, shading = "lift", main = "Scatter Plot of Three Rules Involving Coffee")

# Generate a scatter plot of the same three rules, but with method="graph" and engine="htmlwidget"
plot(pastry_rules, method = "graph", engine = "htmlwidget")
```

"**Task 2: Hierarchical Clustering**"
1. Read the dataset tiktok_top_1000.csv into your R environment. What are your dataset’s dimensions?

```{r}
tiktok <- read.csv('tiktok_top_1000.csv')
dim(tiktok)
```

```{r}
set.seed(14308752)  # set your BUID as the seed
sample_rows <- sample(nrow(tiktok), 25, replace = FALSE)
sampled_data <- tiktok[sample_rows, ]
```

3. Should this data be scaled? Why or why not? If so, scale your data’s numeric variables.

Answer: the numeric variables in this data Subscribers.count, Views.avg., Likes.avg., Comments.avg., Shares.avg should be scaled because they are measured on different scales and have different units of measurement

```{r}
numeric_vars <- c("Subscribers.count", "Views.avg.", "Likes.avg.", "Comments.avg.", "Shares.avg.")
scaled_df <- sampled_data
scaled_df[numeric_vars] <- scale(sampled_data[numeric_vars])
```

4. Build a hierarchical clustering model for the dataset, using any method for inter-cluster
dissimilarity
a. Create and display a dendrogram for your model. By looking at your dendrogram,
how many clusters do you see here? (There is not a single correct answer to this
question, just describe the number of clusters that seem to be showing here).

Answer: I see 5 clusters in Denrogram

```{r}
dist_matrix <- dist(scaled_df[, c("Subscribers.count", "Views.avg.", "Likes.avg.", "Comments.avg.", "Shares.avg.")])

hclust_model <- hclust(dist_matrix, method = "complete")

plot(hclust_model, hang = -1, cex = 0.6, main = "Dendrogram of 25 TikTok accounts")
```
b. Use the cutree function to cut the records into clusters. Specify your desired
number of clusters, and show the resulting assignments for each TikTok artist.

Answer: As dendogrman showed 5 cluster, i picked 5 clusters.

```{r}
cluster_assignments <- cutree(hclust_model, k = 5)

library(dplyr)
clustered_df <- scaled_df %>% mutate(cluster = cluster_assignments)
```
c.Attach the assigned cluster numbers back to the original dataset. Use groupby()
and summarize() from tidyverse to generate per-cluster summary stats, and write 2-3 sentences about what you find. What stands out here? What do you notice about any unusual variables or clusters?

Answer: The summary stats show the average values of subscribers, views, likes, comments, and shares for each cluster. The values range from negative to positive, with higher values indicating greater popularity.

One cluster (cluster 3) stands out with particularly high average values for views, likes, comments, and shares. This suggests that the creators in this cluster have a large following and high engagement with their audience.

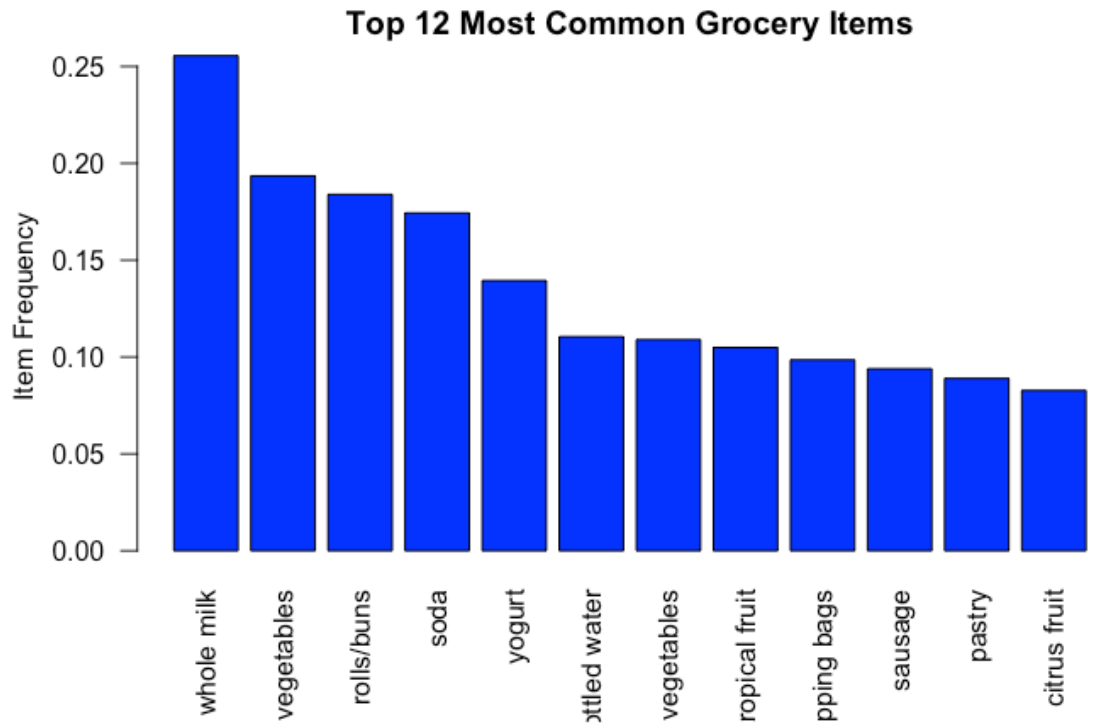
Cluster 4 also stands out with extremely high values for subscribers, indicating that the creators in this cluster have a very large following. Cluster 5 has a negative value for likes, indicating that the creators in this cluster have relatively low engagement with their audience. Additionally, the average value for shares is unusually high, which may suggest that the creators in this cluster rely on shareability rather than engagement to gain visibility on the platform.

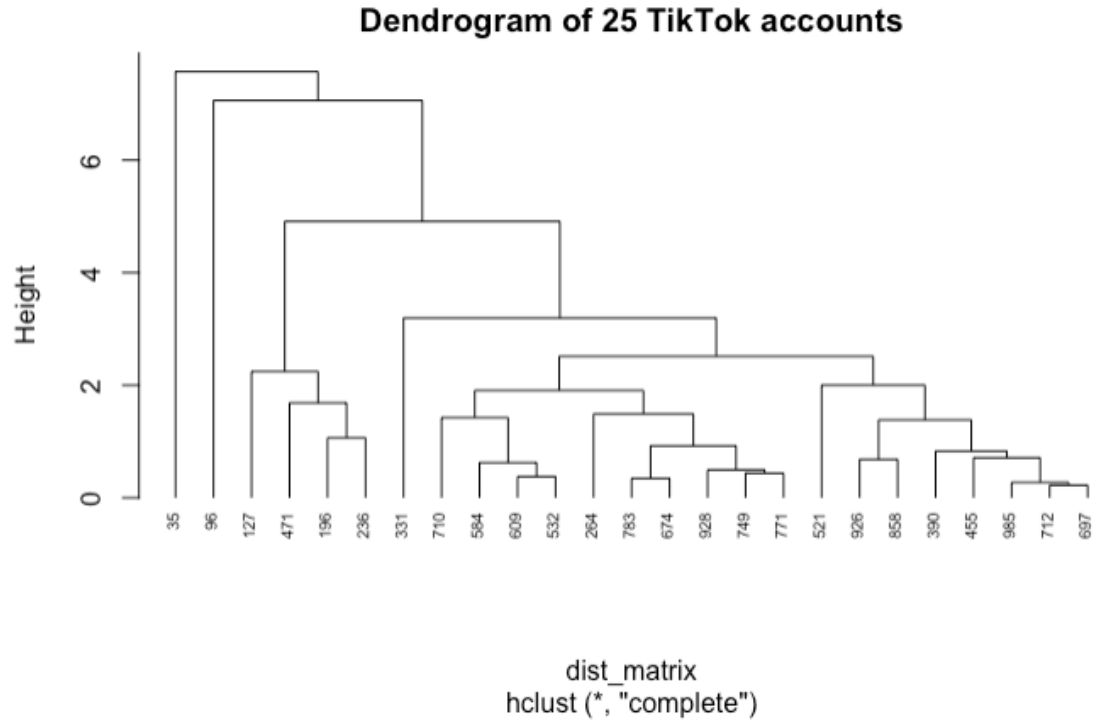
```{r}
summary_stats <- clustered_df %>%
  group_by(cluster) %>%
  summarize(avg_subs = mean(Subscribers.count),
            avg_views = mean(Views.avg.),
            avg_likes = mean(Likes.avg.),
            avg_comments = mean(Comments.avg.),
            avg_shares = mean(Shares.avg.)) %>%
  arrange(cluster)
head(clustered_df)
```

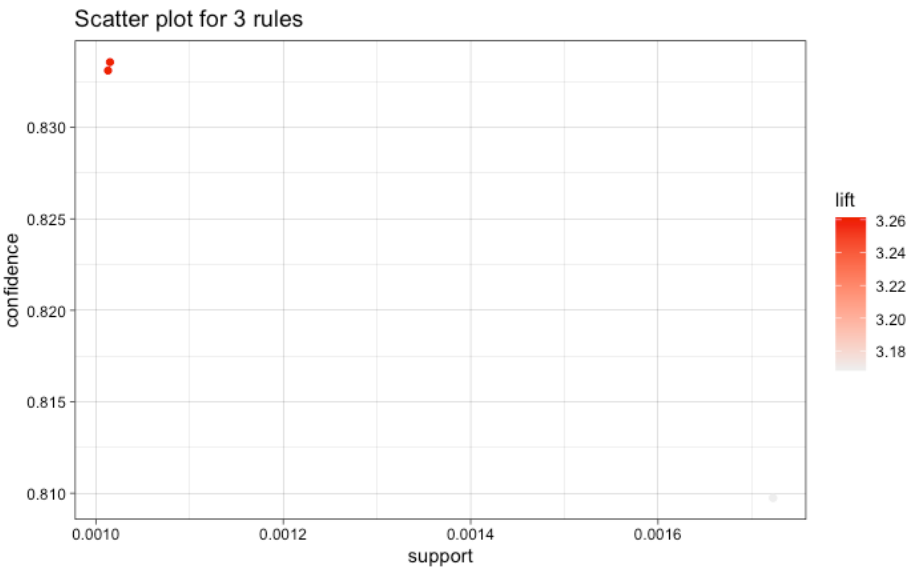
5. Choose any TikTok artist from among your 25. What cluster did it fall into? Write 2-3
sentences about the other members of its cluster (or if it’s a singleton, write a bit about
why it is a singleton).

Answer:

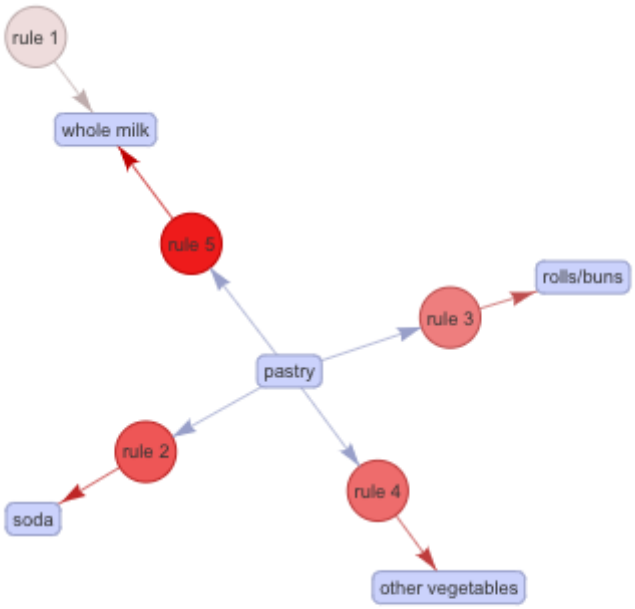
Samseats falls into cluster 2. This cluster is characterized by having relatively high views, likes, and shares, but low subscriber counts. Other members of this cluster include Madison Pettis and Frankie Fictitious. These TikTok artists may have gained popularity for their content or style of videos, but they may not have a dedicated fan base or consistent viewership.


Top 12 Most Common Grocery Items
Item Frequency
0.25
0.20
0.15
0.10
0.05
0.00
whole milk
vegetables
rolls/buns
soda
yogurt
ottled water
vegetables
ropical fruit
pping bags
sausage
pastry
citrus fruit


Dendrogram of 25 TikTok accounts
Height
0
2
4
6
35
96
127
471
196
236
331
710
584
609
532
264
783
674
928
749
771
521
926
858
390
455
985
712
697
dist_matrix
hclust (*, "complete")


Scatter plot for 3 rules
confidence
0.830
0.825
0.820
0.815
0.810
0.0010
0.0012
0.0014
0.0016
support
lift
3.26
3.24
3.22
3.20
3.18

Select by id


rule 1
whole milk
rule 5
rolls/buns
rule 3
pastry
rule 2
rule 4
soda
other vegetables