Cluster_Cocacola & Dr Pepper

Stephanie Yang jy2777

12/2/2017

Coca Cola #4 Special

```
setwd("~/Desktop/Homework/Statistical Methods/Project/datasets")
#transpose t()
cocacola.metrics <- t(read.csv("cocacola_metrics2.csv", header=FALSE))

#labels columns
colnames(cocacola.metrics)<- c("Date","Likes (Total) FB","Comments (Total) FB","Shares (Total) FB","Rea

#removes duplicate row
cocacola.metrics1 <- cocacola.metrics[-1,]

##Cleaning the metrics sheet
#1) removes space in column titles
colnames(cocacola.metrics1) <- gsub(" ","",colnames(cocacola.metrics1))

#2) removes % symbol of column 10
cocacola.metrics1[,c(11,15,19,20,21,26,27,28,31,33,38,39,43,44,52,59,60,61,69,71,72,73)] <- as.numeric(
cocacola.metrics1 <- as.data.frame(cocacola.metrics1)
class(cocacola.metrics1)
```

```
## [1] "data.frame"
```

```
#3) removes comma separator for thousands, except for date column which is type character not numeric
#gsub to replace "," with "", and then convert the string to numeric using as.numeric
cocacola.metrics1[,2:73] <- lapply(cocacola.metrics1[,2:73], function(x) as.numeric(gsub(",","", as.cha
```

```
## Warning in FUN(X[[i]], ...): NAs introduced by coercion
```

```
##Transforms Monthly to Quarterly Data:
library("lubridate")
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
#creates a year and quarter column per row
cocacola.metrics1$Date <- ymd(cocacola.metrics1$Date)
```

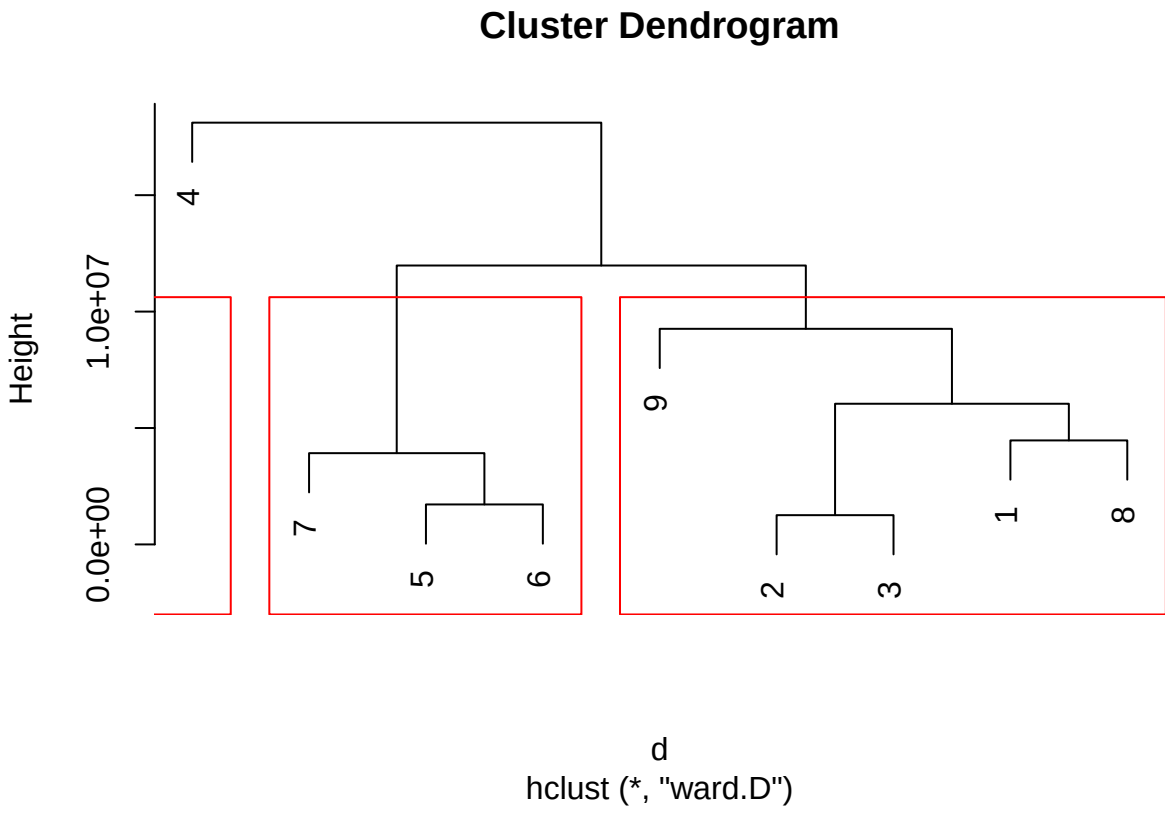
```
## Warning in as.POSIXlt.POSIXct(x, tz): unknown timezone 'zone/tz/2017c.1.0/
## zoneinfo/America/New_York'
```

```
cocacola.metrics1$year = year(cocacola.metrics1$Date)
cocacola.metrics1$quarter = quarter(cocacola.metrics1$Date)

#aggregates quarters of same year and takes their sum (sales are aldo sums) : HOW TO DO IT WITH AGGREGA
library("reshape2")
cocacola.metrics2 <- melt(cocacola.metrics1[,2:75], id=c("quarter", "year"))
```

```
cocacola.metrics2 <- dcast(cocacola.metrics2, year + quarter ~ variable, fun.aggregate = sum)
write.csv(cocacola.metrics2,file="cocacola_vizmetrics.csv")

d <- dist(cocacola.metrics2, method = "euclidean") # distance matrix
fit <- hclust(d, method="ward")
```

```
## The "ward" method has been renamed to "ward.D"; note new "ward.D2"
```

```
plot(fit) # display dendogram
groups <- cutree(fit, k=3) # cut tree into 3 clusters
# draw dendogram with red borders around the 3 clusters
rect.hclust(fit, k=3, border="red")
```



Dr. Pepper

```
setwd("~/Desktop/Homework/Statistical Methods/Project/datasets")
#transpose t()
drpepper.metrics <- t(read.csv("drpepper_metrics2.csv", header=FALSE))

#labels columns
colnames(drpepper.metrics)<- c("Date","Likes (Total) FB","Comments (Total) FB","Shares (Total) FB","Rea

#removes duplicate row
drpepper.metrics1 <- drpepper.metrics[-1,]

##Cleaning the metrics sheet
#1) removes space in column titles
colnames(drpepper.metrics1) <- gsub(" ","",colnames(drpepper.metrics1))
```

```
#2) removes % symbol of column 10
drpepper.metrics1[,c(11,15,19,20,21,26,27,28,31,33,38,39,43,44,52,59,60,61,69,71,72,73)] <- as.numeric(g
drpepper.metrics1 <- as.data.frame(drpepper.metrics1)
class(drpepper.metrics1)
```

```
## [1] "data.frame"
```

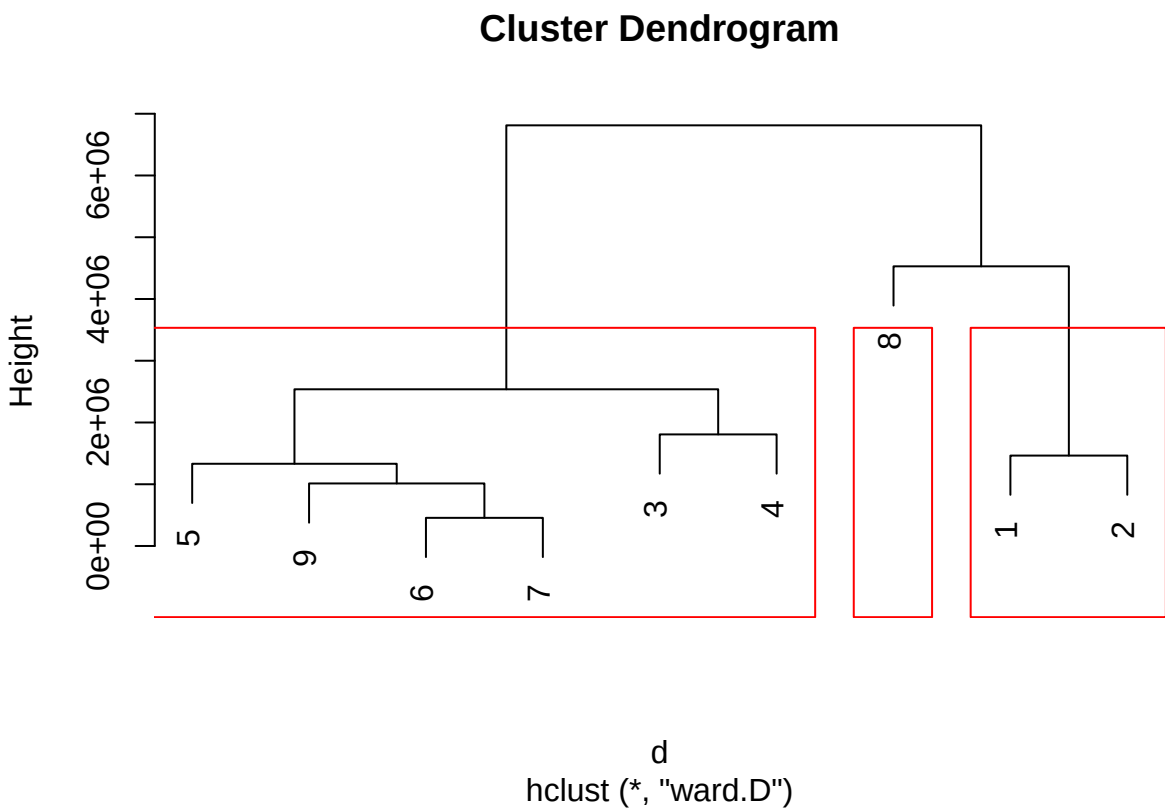
```
#3) removes comma separator for thousands, except for date column which is type character not numeric
#gsub to replace "," with "", and then convert the string to numeric using as.numeric
drpepper.metrics1[,2:73] <- lapply(drpepper.metrics1[,2:73], function(x) as.numeric(gsub(",","", as.cha
```

```
## Warning in FUN(X[[i]], ...): NAs introduced by coercion
```

```
##Transforms Monthly to Quarterly Data:
library("lubridate")

#creates a year and quarter column per row
drpepper.metrics1$Date <- ymd(drpepper.metrics1$Date)
drpepper.metrics1$year = year(drpepper.metrics1$Date)
drpepper.metrics1$quarter = quarter(drpepper.metrics1$Date)

#aggregates quarters of same year and takes their sum (sales are aldo sums) : HOW TO DO IT WITH AGGREGA
library("reshape2")
drpepper.metrics2 <- melt(drpepper.metrics1[,2:75], id=c("quarter", "year"))
drpepper.metrics2 <- dcast(drpepper.metrics2, year + quarter ~ variable, fun.aggregate = sum)
write.csv(drpepper.metrics2,file="cocacola_vizmetrics.csv")

d <- dist(drpepper.metrics2, method = "euclidean") # distance matrix
fit <- hclust(d, method="ward")
```

```
## The "ward" method has been renamed to "ward.D"; note new "ward.D2"
```

```
plot(fit) # display dendogram
groups <- cutree(fit, k=3) # cut tree into 3 clusters
# draw dendogram with red borders around the 3 clusters
rect.hclust(fit, k=3, border="red")
```


Cluster Dendrogram
Height
0e+00
2e+06
4e+06
6e+06
5
9
6
7
3
4
8
1
2
d
hclust (*, "ward.D")