Using R for Statistical Analysis of Survey Data

javaria

20 Dec 2020

Herrarchial Clustering Using R

Intro

R is a natural place to play with data and collections of natural / artificial data recorders, commonly in the form of electronic devices, ranging fromm cell phones to embedded platforms.

Process

1. Install R
2. Prepare csv from input file (Excel .xlsx)
3. Start R

Assignment

2. Excel Sheet to be downloaded from link provided
3. Perform MBA Using R and share most watched
4. Create appropriate visualizations to describe variables and answer the following questions :
    - Find top 3 movies?
    - Which age range likes superhero movies?
5. Using Hierarchial Clustering to find which gender / age group likes which of the top favourite movies?

Solution

Loading Excel Data

First Excel file needs to be converted to a CSV file by going to `File -> Save As` and selecting CSV (comma seperated values) as the format. Now it's time to load up R from any console and get to the R shell. The csv file can be loaded by calling `read.csv()` function

```
DF<-read.csv('input.csv',sep=';')
head(DF)
```

```
##             Timestamp Age Gender favourite medium
## 1 11/27/2020 11:15:59  21
## 2 11/27/2020 11:16:07  21
## 3 11/27/2020 11:17:44  21
## 4 11/27/2020 11:19:07  24
## 5 11/27/2020 11:19:14  23
## 6 11/27/2020 11:21:03  22
##
list..
## 1                                                                      Fast and Furio
us, Captain America, Avenger - End Game, Top Gun, Interstellar, Pirates of the Caribbea
n, Five Feet Apart, The Fault in Our Stars, The Dark Knight, Joker, The Pursuit of Happi
ness, Legend, Titanic, The Godfather, Mission Impossible, Matrix Reloaded, Jurassic Par
k, Deep Blue Sea, Home Alone, Final Destination, 127 hours, Godzilla, Shawshank Redempti
on, The good, the bad, the ugly, The Wolf of Wall Street, Charlie and the Chocolate Fact
ory
## 2
Fast and Furious, Gravity, Captain America, Avenger - End Game, Top Gun, Pirates of the
Caribbean, The Dark Knight, Joker, Legend, Titanic, The Godfather, Mission Impossible, L
ord of the Rings, Matrix Reloaded, Jurassic Park, Home Alone, Jumanji, Final Destinatio
n, 127 hours, Godzilla, Shawshank Redemption, Once upon a time in Mexico, The good, the
bad, the ugly, The Wolf of Wall Street, Charlie and the Chocolate Factory
## 3
Fast and Furious, Titanic, Home Alone, Charlie and the Chocolate Factory
## 4 Contagion, Fast and Furious, Gravity, Captain America, Avenger - End Game, Top Gun,
Interstellar, Pirates of the Caribbean, The Fault in Our Stars, The Dark Knight, Joker,
The Pursuit of Happiness, Legend, Titanic, The Godfather, 7 pounds, Mission Impossible,
Lord of the Rings, Matrix Reloaded, Jurassic Park, Deep Blue Sea, Home Alone, Jumanji, F
inal Destination, 127 hours, Godzilla, Shawshank Redemption, Once upon a time in Mexico,
The good, the bad, the ugly, The Wolf of Wall Street, Charlie and the Chocolate Factory
## 5
Fast and Furious, Captain America, Avenger - End Game, Top Gun, Pirates of the Caribbea
n, Five Feet Apart, The Fault in Our Stars, The Dark Knight, The Godfather, Mission Impo
ssible, Lord of the Rings, Jurassic Park, Deep Blue Sea, Home Alone, Jumanji, Final Dest
ination, Godzilla, The Wolf of Wall Street, Charlie and the Chocolate Factory
## 6
Fast and Furious, Captain America, Avenger - End Game, Interstellar, Pirates of the Cari
bbean, The Dark Knight, Titanic, Lord of the Rings, Jurassic Park, Shawshank Redemption,
The Wolf of Wall Street
```

The data from the file is now in the variable `DF` and the column names can be seen by calling `colnames()` function

```
print(colnames(DF))
```

```
## [1] "Timestamp" "Age"       "Gender"    "favourite" "medium"    "list.."
```

Data needs some cleaning up

```
DF$list..<-NULL
DF<-DF[40:189,]
DF$favourite<-tolower(trimws(DF$favourite))
```

Perform MBA Using R and share most watched

By using builtin summary of the column "favourite movie" we find that the top 10 movies are :

```
head(summary(DF$favourite),n=10)
```

```
##    Length     Class      Mode 
##       150 character character
```

Visualizations to describe variables to answer the following :

- Find top 3 movies?

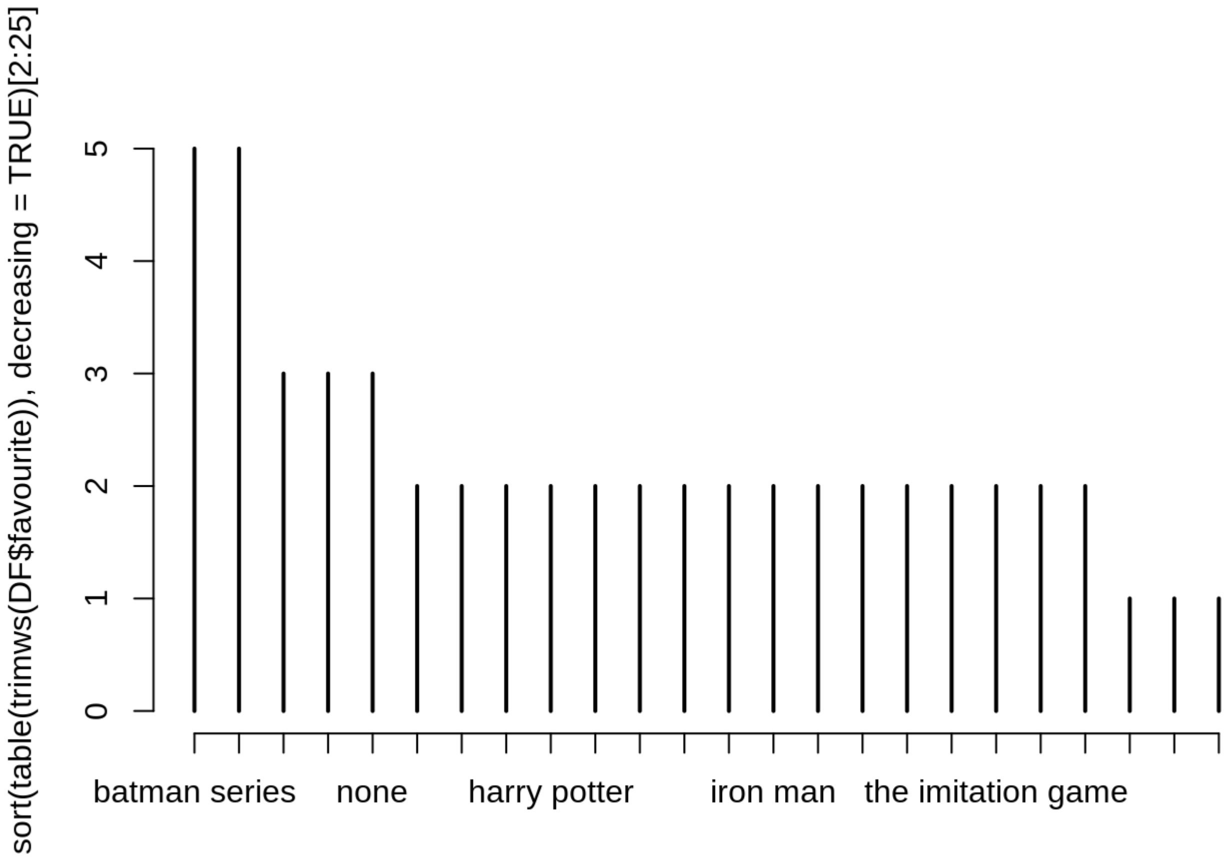
```
head(summary(DF$favourite),n=4)
```

```
##    Length     Class      Mode 
##       150 character character
```

```
#plot(head(summary(DF$favourite),n=4))
```

Ignoring the first entry as it contains empty spaces in the survey dataset

```
plot(sort(table(trimws(DF$favourite)),decreasing=TRUE)[2:25])#plots after doing all othe
r stuff
```



- Which age range likes superhero movies?

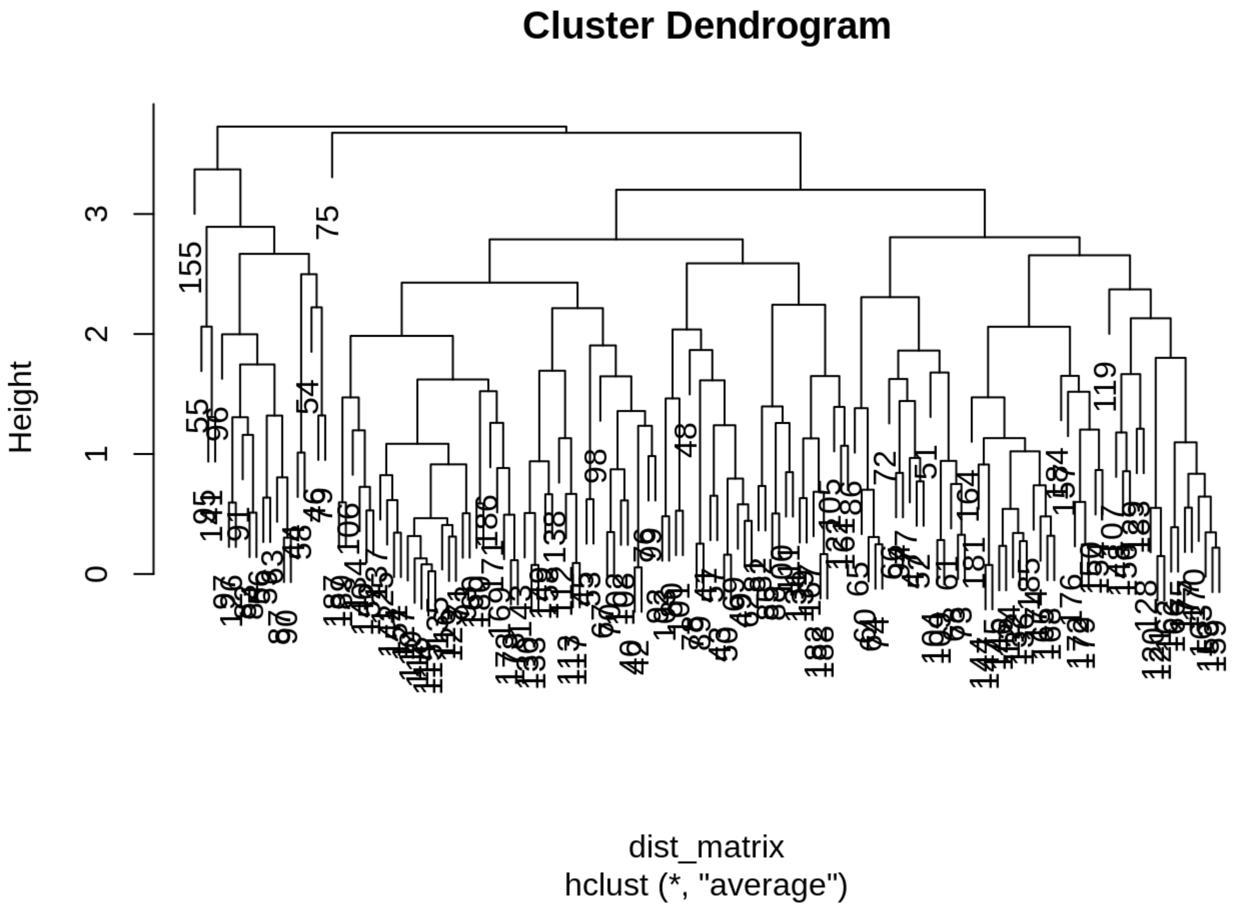
```
superheroes=c("Joker","Batman","Avenger","Captain")
superherofans=NULL
for (hero in superheroes){
superherofans<-append(superherofans,subset(DF, grepl(hero,favourite)))
}
#superherofans
```

We have seperated all superhero fans in a seperate list now

Using Hierarchial Clustering to find which gender / age group likes which of the top favourite movies?

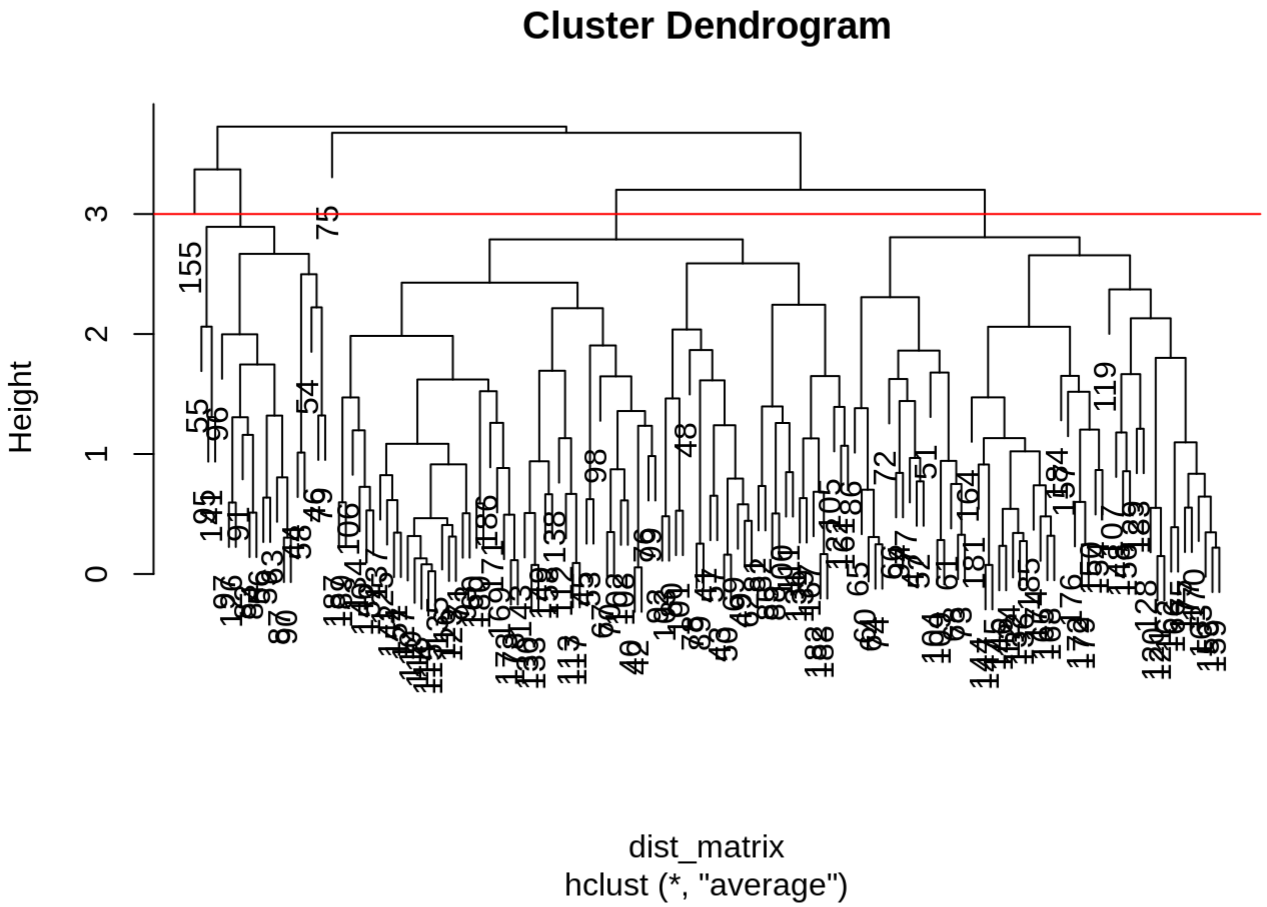
```
for (col in 1:ncol(DF)){
DF[,col]<-as.integer(as.factor(DF[,col]))
DF[,col]<-scale(DF[,col])
}
```

```
dist_matrix=dist(DF,method='euclidean')
hclust_avg=hclust(dist_matrix,method='average')
plot(hclust_avg)
```

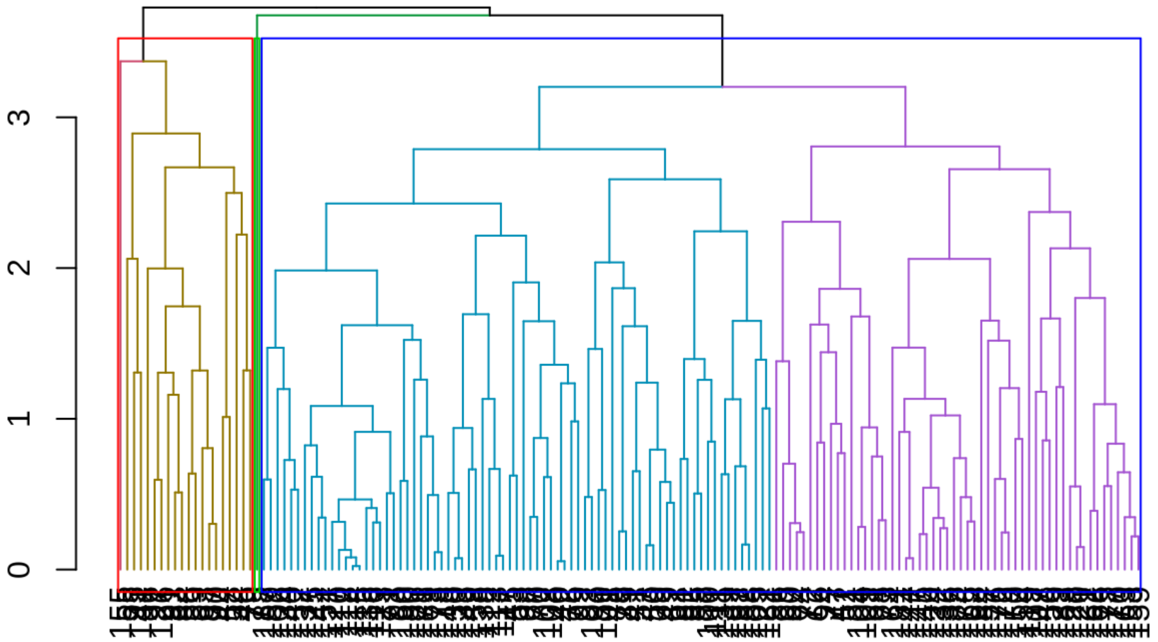


Now time to cut the tree on 3rd level

```
for (col in 1:ncol(DF)){
DF[,col]<-as.integer(as.factor(DF[,col]))
DF[,col]<-scale(DF[,col])
}
dist_matrix=dist(DF,method='euclidean')
hclust_avg=hclust(dist_matrix,method='average')
plot(hclust_avg)
#plot.new()
cut_avg=cutree(hclust_avg,k=3)
abline(h=3,col='red')
avg_dend_obj=as.dendrogram(hclust_avg)
suppressPackageStartupMessages(library(dendextend))
```



```
avg_col_dend=color_branches(avg_dend_obj,h=3)
plot(avg_col_dend)
rect.hclust(hclust_avg,k=3,border=2:6)
```


0
1
2
3