Comedy Service Case

Mitchell J. Lovett

11/14/2020

R Markdown

```
#install.packages("cluster")
#install.packages("fpc")
library(cluster)
library(fpc)
```

```
## Warning: package 'fpc' was built under R version 3.6.2
```

```
library(foreign)

## Post Hoc Segmentation (cluster analysis)
dir = "~/Dropbox/Analytics Design/Cases/Comedy Study/"
setwd(dir)
filnm = "comedy"
spssdatalab <- read.spss(paste(filnm,".sav",sep=""),to.data.frame=TRUE,use.value.labels=TRUE,trim_value
spssdata <- read.spss(paste(filnm,".sav",sep=""),to.data.frame=TRUE,use.value.labels=FALSE)
attr(spssdata, "variable.labels")
```

```
##                                                                 Brian
##                                                           "Brian Regan"
##                                                                  Jeff
##                                                            "Jeff Allen"
##                                                                Stewart
##                                                           "Stewart Lee"
##                                                                 Lewis
##                                                           "Lewis Black"
##                                                                 Maria
##                                                         "Maria Bamford"
##                                                                   Jim
##                                                            "Jim Gaffigan"
##                                                                Russell
##                                                          "Russell Peters"
##                                                                  Dave
##                                                         "Dave Chappelle"
##                                                                 Pablo
##                                                       "Pablo Francisco"
##                                                                  Nick
##                                                          "Nick Swardson"
##                                                              ComedyAtt_1
##                                  "I was familiar with the comedians shown"
##                                                              ComedyAtt_2
##                                       "Overall I found these clips funny"
```

```
##                                                                 ComedyAtt_3
##                                "Watching stand up comedy is important to me"
##                                                                 ComedyAtt_4
##                               "I consider myself knowledgeable about comedy"
##                                                                 ComedyAtt_5
##                           "Compared to most people I know a lot about comedy"
##                                                                 ComedyAtt_6
##                                           "I watch a lot of stand up comedy"
##                                                                 ComedyAtt_7
## "Prior to this service, I had previously seen some of the comedy clips shown"
##                                                                 ComedyAtt_8
##                                        "I watch videos on youtube regularly"
##                                                                 ComedyAtt_9
##  "I personally felt the language in some of the clips I viewed was offensive"
##                                                                         Sex
##                                                       "What is your gender?"
##                                                                         Age
##                                                          "How old are you?"
##                                                                        Race
##                                                        "What is your race?"
##                                                                         Edu
##                  "What is the highest level of education you have completed?"
```

Above loads the data for the comedy service case study. This data is from a test of a comedy service where you subscribe to receive short clips of the best comedy clips on the web curated for you. The test was trying to evaluate whether there are different tastes for comedians across individuals (to support potential targeted offers). This is so-called horizontal differentiation in that one person might like a set of comedians than the tastes of other people.

To evaluate this, we will conduct a segmentation analysis using cluster analysis. The data contains liking for ten comedians along with responses to some comedy attitudes. All but one scale has higher being better. The last scale is about the use of offensive language in comedy and is reverse scored (higher values means likes comedy with offensive language less).

For the analysis we will construct a subset of the data for the cluster analysis. This will focus us on the data we care about. We then use `kmeans()` to implement the cluster analysis.

K-means clustering via `kmeans` takes as arguments data and the number of clusters. It constructs clusters that try to minimize the distance within cluster and increase the distance between clusters (homogeneity within and heterogeneity between). Because k-means clustering assumes the data has distance it should only be applied to interval or ratio measures, or two category nominal measures (e.g., 0 and 1 values) such as dummy variables.

K-means clustering is not deterministic and depends on the initial conditions. Hence, we normally want to set the random seed via `set.seed()` prior to calling the routine. This allows us to reproduce the results in the future by calling the same seed.

We will first use a function clustTest in the file ClusterCode.R that tests how many clusters to use. This code takes a number of optional arguments, and you must minimally pass the data. That code automatically sets the seed, runs multiple starting points for the cluster analysis across a range of numbers of clusters, and produces figures containing visualizations of the measures of the quality of fit.

The first measure is the within sum of squared errors. This is the errors calculated as the distance between the cluster centroids (middle) and the data points assigned to that cluster. These errors are similar to the residuals in a linear regression and this is similar to the sum of squared errors in the linear regression context. To identify the best number of clusters in this plot we follow the elbow rule. We look for where the plot seems the bend sharply to be flatter and normally consider that point and one point just before or after that point.

The second measure is the average silhouette width. This measures both the homogeneity within and the heterogeneity between. We want a higher value of the average silhoutte width, since that corresponds with a better quality fit. See https://en.wikipedia.org/wiki/Silhouette_(clustering) for more details.

Normally, I consider both approaches and use both to inform the number of clusters to consider for more detailed interpretation.

```
set.seed(123456)   # set random number seed before doing cluster analysis
toClust = spssdata[,1:19]    # select the relevant data for clustering
source("ClusterCode.R")
```

```
## Loading required package: factoextra
```

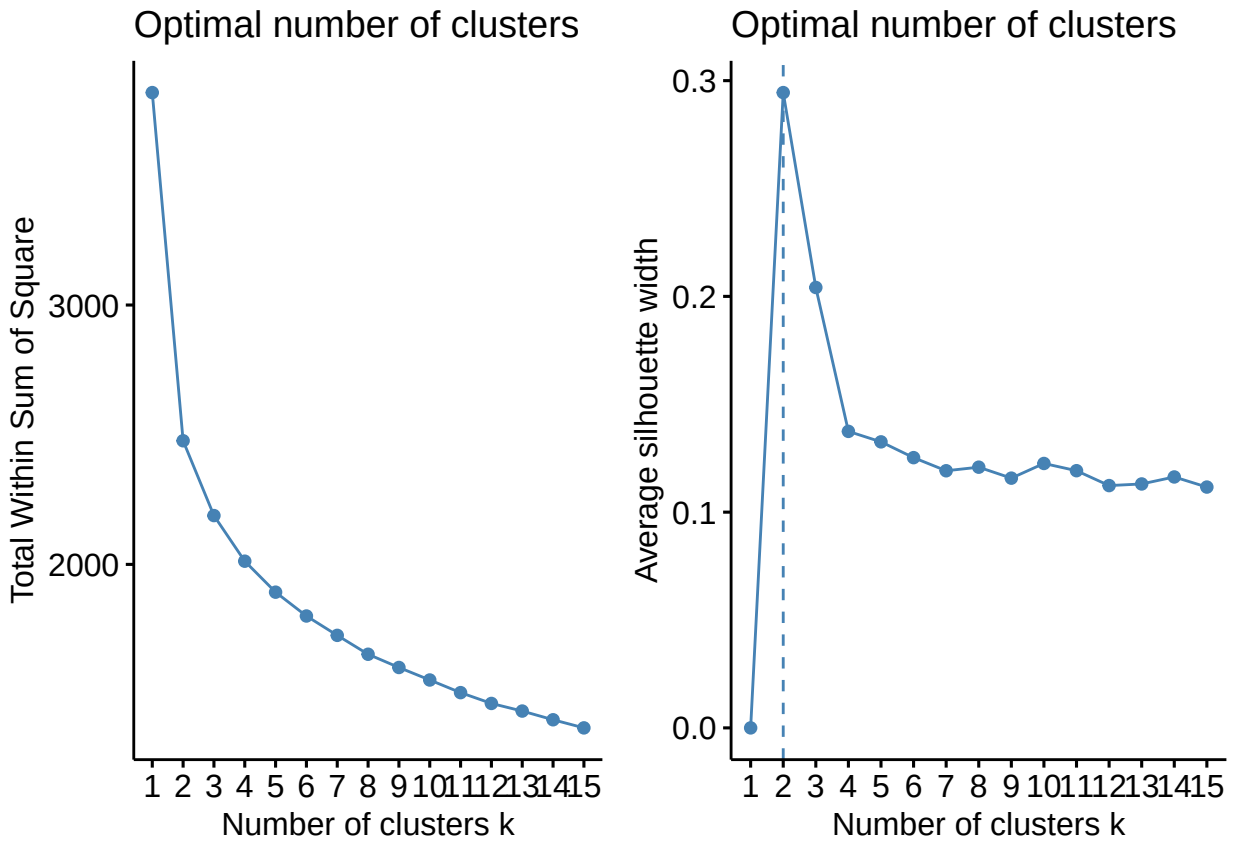
```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.6.2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

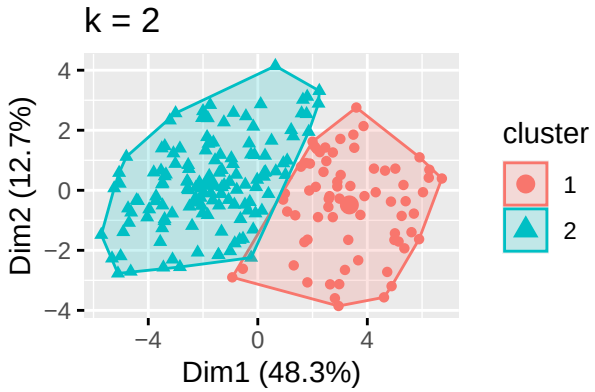
```
## Loading required package: gridExtra
```

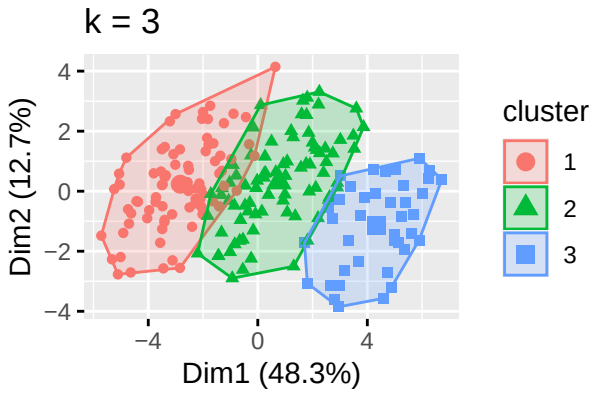
```
tmp = clustTest(toClust)
```



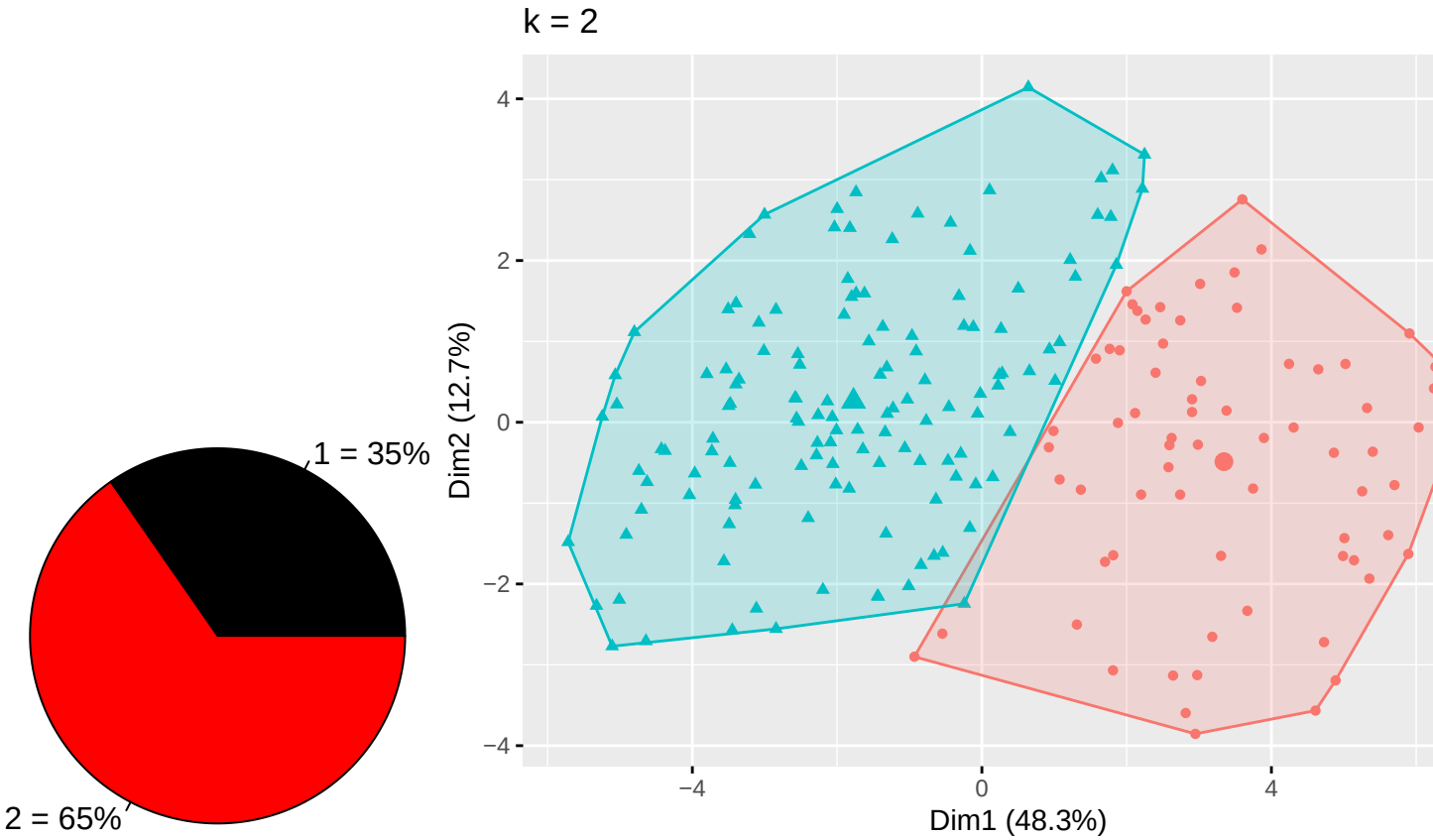
Here we find by the elbow rule that two or three clusters looks best, though the elbow is not very distinct. The optimal number of clusters by the average silhouette width is clearly two. As a result, we will focus on interpreting two and three cluster solutions.

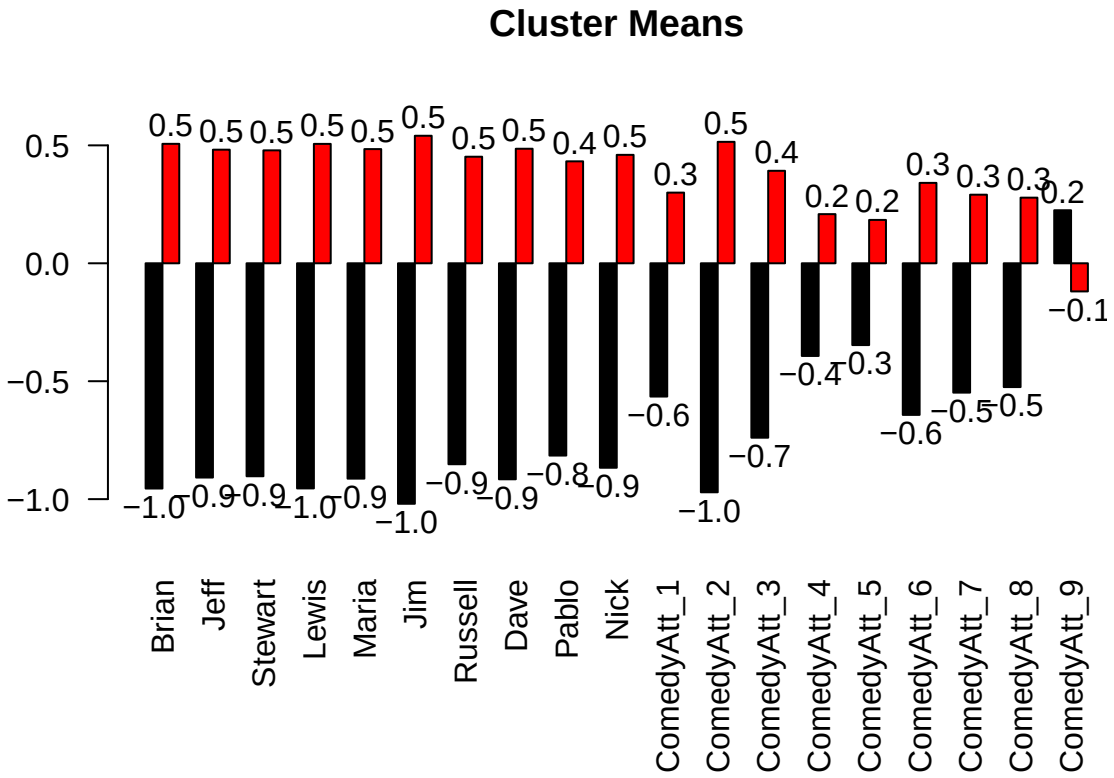
```
clusts = runClusts(toClust,2:3)
```



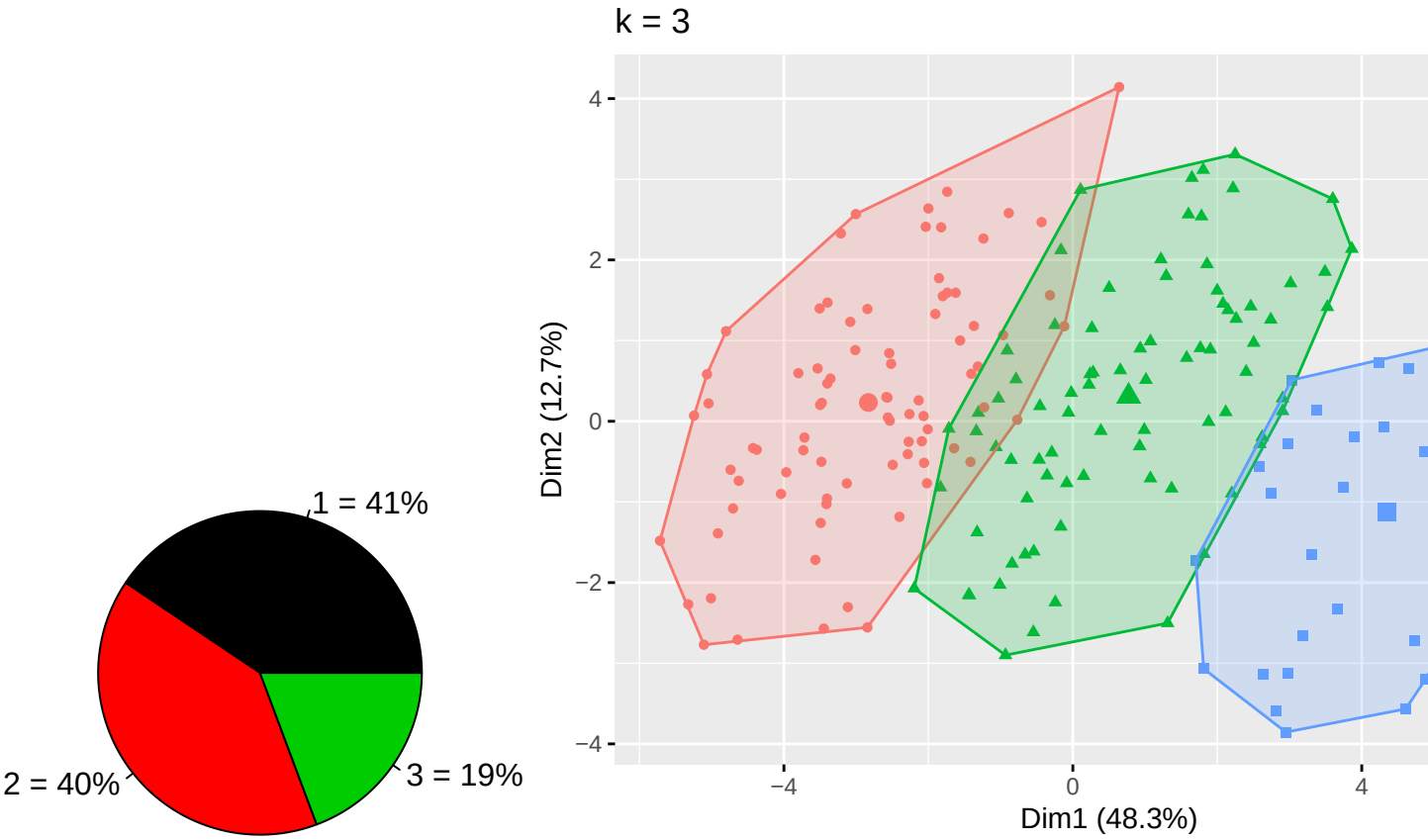


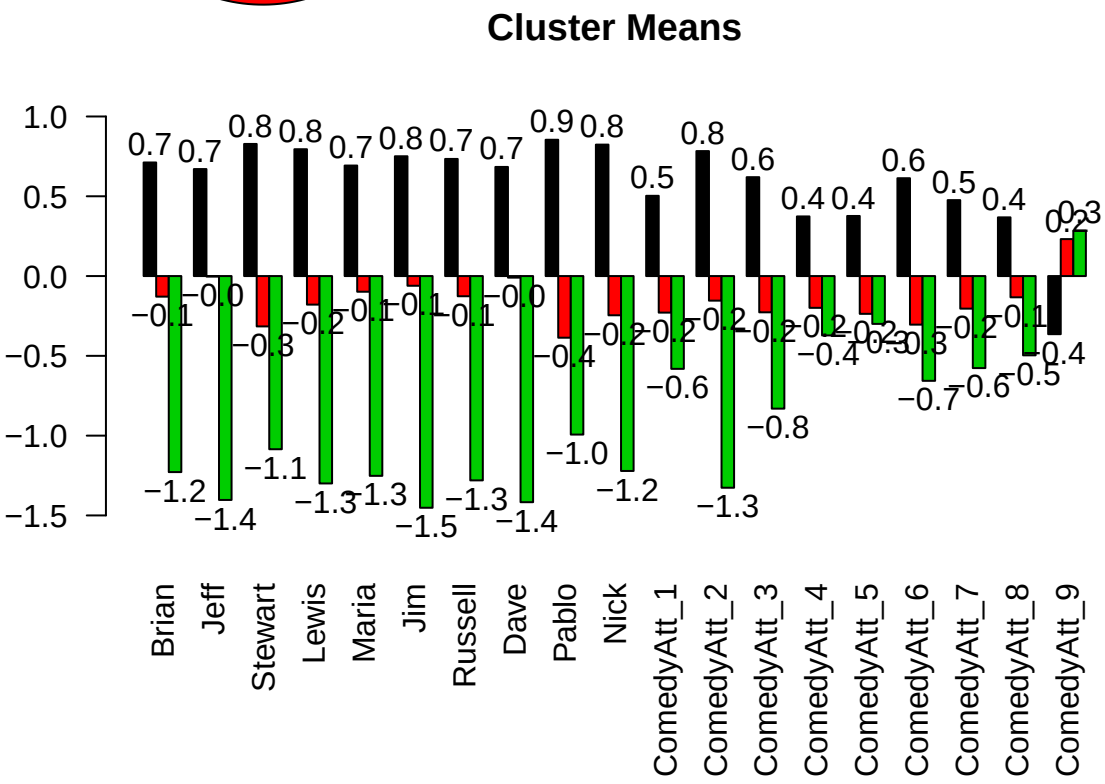

```
plotClust(clusts$kms[[1]],toClust)
```






```
plotClust(clusts$kms[[2]],toClust)
```





We find that the two cluster solution has a 35-65% split. The segments are quite distinct and separated through both of the first two principle components. This implies that many variables are involved in separating the clusters. The cluster means barplot indicates that the larger segment is more positive about comedy in

general and that there is no horizontal differentiation in tastes.

The three cluster solution splits the sample into three segments having 41-19-40%, corresponding to black, green, and red. The segments overlap more visually, and this is only for the two principle components. Like in the two segment model, the cluster analysis distinguishes on both of the first two principle components.Again the segmentation largely separates the customers into vertical groups who prefer all comedy more or less.

Comparing the two and three segment versions reveals that little additional insight about *horizontal* differentiation is available from the more complicated three segment version. Hence, we would likely go with the two segment model that is preferred by the elbow rule and silhouette measure.